SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1997

•

VOLUME 23

•

NUMBER 1

Statistics Canada Statistique Canada

Canada

FIRST CALL FOR PAPERS

WORKSHOP AND SYMPOSIUM ON LONGITUDINAL ANALYSIS FOR COMPLEX SURVEYS

Statistics Canada, Ottawa, Canada
May 19-22, 1998

Statistics Canada's *XVth* annual international methodology symposium will be on the topic of longitudinal analysis for complex surveys. In conjunction with this symposium, *Statistics Canada* and the *Centre de recherches mathématiques* (*CRM*), *Université de Montréal*, are sponsoring a workshop on this same topic. This workshop is one of the many events taking place during the *CRM*'s theme year in statistics.

The focus of Symposium '98 is on recently developed methods in longitudinal data analysis. Emphasis will be given to the theory and application of longitudinal methods for data from complex surveys. The symposium will give participants an opportunity to meet colleagues who are involved in solving problems unique to the analysis of survey data, including *David Binder, Wayne Fuller, Harvey Goldstein, Lisa Lavange, Jerry Lawless, Danny Pfeffermann, and J.N.K. Rao.*

We invite abstracts for papers related to the theme of Symposium '98. A non-exhaustive list of topics is included with this invitation. Papers concerning new or previously undocumented approaches, methodologies and applications are especially welcome. Academic researchers and practitioners from both the private and public sectors are encouraged to submit.

Abstracts of 200-300 words, in English or French, along with the presenter's name, affiliation, complete address, telephone and fax numbers and email address, should be sent to the address below. **The deadline for abstracts is October 31, 1997. The final selection of papers will be announced by December 31, 1997.**

Submit abstracts to:

Michael Hidiroglou
Statistics Canada
11th floor, R.H. Coats Building
Ottawa, Ontario
Canada K1A 0T6
Telephone: (613)951-4767
Fax: (613)951-1462
email: symposium98@statcan.ca

Presenters must submit a draft paper, in English or French, by April 17, 1998, for the purposes of official simultaneous translation. The final version of a paper must be provided by June 30, 1998, in order to appear in the symposium proceedings.

<u>**Non-exhaustive list of topics:**</u>

Preparing/storing survey data for longitudinal analysis.

Imputing for longitudinal data analysis.

Weighting issues with longitudinal surveys.

Gross flows - Methods of estimation and applications.

Multi-level modelling techniques and applications to longitudinal survey data (including random effects models).

Event history techniques and applications with survey data.

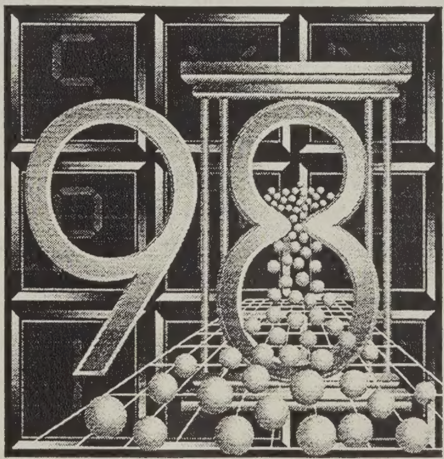
Marginal modelling and applications with survey data.

Software for applying longitudinal techniques to survey data.

Causal analysis of panel data.

For more information, please visit our web site:

www.statcan.ca/english/conferences/symposium98/index.htm



PREMIER APPEL D'ARTICLES

ATELIER ET SYMPOSIUM SUR L'ANALYSE LONGITUDINALE POUR LES ENQUÊTES COMPLEXES

Statistique Canada, Ottawa, Canada
19-22 mai 1998

Le *XV*[e] symposium international méthodologique annuel de *Statistique Canada* portera sur l'analyse longitudinale pour des enquêtes complexes. Dans le cadre du symposium et en collaboration avec le *Centre de recherches mathématiques (CRM)* de *l'Université de Montréal*, *Statistique Canada* tiendra un atelier sur le même sujet. Le thème du *CRM* de cette année est la statistique et l'atelier est un des nombreux événements qui se dérouleront en 1998.

Le centre d'intérêt du symposium '98 sera les développements récents dans l'analyse de données longitudinales. L'emphase sera mise sur la théorie et les applications dans le cas de données provenant d'enquêtes avec un plan de sondage complexe. Le symposium donnera aux participants une opportunité de rencontrer des collègues impliqués dans la résolution de problèmes liés aux enquêtes tels *David Binder, Wayne Fuller, Harvey Goldstein, Lisa Lavange, Jerry Lawless, Danny Pfeffermann, et J.N.K. Rao.*

Nous vous invitons à soumettre des résumés de communications portant sur le thème du symposium '98. Une liste non-exhaustive de sujets est incluse avec cette invitation. Des articles sur des nouvelles approches ou des approches non documentées jusqu'à maintenant sont particulièrement bienvenus. On encourage les chercheurs universitaires et les divers intervenants des secteurs public et privé à soumettre leurs communications.

Un résumé de 200 à 300 mots devrait être envoyé, en français ou en anglais, à l'adresse qui suit, accompagné du nom de la personne qui présentera la communication, de son affiliation, de son adresse complète, de ses numéros de téléphone et de télécopieur, et de son adresse électronique. **La date limite pour soumettre un résumé est le 31 octobre 1997. On annoncera les communications libres et invitées qui auront été retenues le 31 décembre 1997.**

Soumettre les résumés à :

Michael Hidiroglou
Statistique Canada
11ième étage, R.H. Coats
Ottawa, Ontario
Canada K1A 0T6
Téléphone : (613)951-4767
Télécopieur : (613)951-1462
courrier électronique : symposium98@statcan.ca

Les conférenciers doivent soumettre une ébauche de la communication, en anglais ou en français, au plus tard le 17 avril 1998, pour permettre la traduction simultanée. Les actes du symposium comprenant les communications libres et invitées seront publiés. Pour être incluse dans les actes du symposium, la version finale de la communication devra être soumise au plus tard le 30 juin 1998.

<u>Liste non-exhaustive de sujets:</u>

Préparation/stockage des données pour l'analyse de données longitudinales.

Imputation pour l'analyse de données longitudinales.

Pondération pour les enquêtes longitudinales.

Flux bruts - Méthodes d'estimation et applications.

Modèles hiérarchiques et applications aux données longitudinales d'enquête (incluant modèles à effets aléatoires).

Modèles de survie et applications avec des données d'enquête.

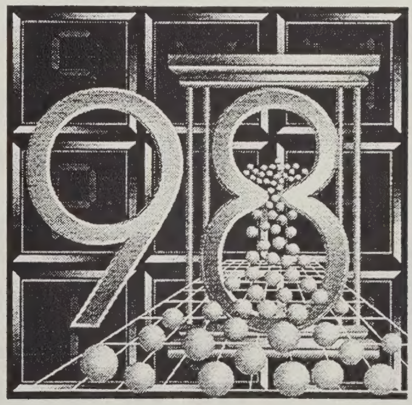
Modèles marginaux et applications aux données d'enquête.

Logiciels pour l'application de techniques longitudinales aux données d'enquête.

Analyse causale de données par panel.

Pour plus d'information, veuillez consulter notre site internet :

www.statcan.ca/francais/conferences/symposium98/index_f.htm



Canada

Statistique Canada Statistics Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1997 • VOLUME 23 • NUMBER 1

Published by authority of the Minister responsible for Statistics Canada

July 1997

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 23, Number 1, June 1997

CONTENTS

In This Issue 1

J.E. STAFFORD and D.R. BELLHOUSE
A Computer Algebra for Sample Survey Theory 3

S. HINKINS, H.L.OH and F. SCHEUREN
Inverse Sampling Design Algorithms 11

P.L.D. NASCIMENTO SILVA and C.J. SKINNER
Variable Selection for Regression Estimation in Finite Populations 23

J.L. ELTINGE and I.S. YANSANEH
Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey 33

M.S. KOVAČEVIĆ and W. YUNG
Variance Estimation for Measures of Income Inequality and Polarization - An Empirical Study 41

K. HUMPHREYS and C.J. SKINNER
Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error 53

J. WAKSBERG, D. JUDKINS and J.T. MASSEY
Geographic-Based Oversampling in Demographic Surveys of the United States 61

W.C. LOSINGER
A Modified Random Groups Standard Error Estimator 73

K. ZEELENBERG
A Simple Derivation of the Linearization of the Regression Estimator 77

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Stafford and Bellhouse, in the first paper, present the basic building blocks to develop a comprehensive computer algebra for survey sampling theory. They show that three basic techniques in sampling theory depend on the repeated application of rules that give rise to partitions. The methodology is illustrated through applications to moment calculation of the sample mean, the ratio estimator and the regression estimator under the special case of simple random sampling without replacement. The machine application to the methodology described was done in the programming language *Mathematica.*

Hinkins, Oh and Scheuren introduce a new strategy for analysis of data from complex surveys. They draw a sub-sample in such a way that the sub-sample may be considered to be a simple random sample from the original population and then apply standard procedures for IID data. They suggest repeating the procedure many times to recover information lost in sub-sampling the original sample. They show how to implement their approach for stratified element sampling, for one and two stage cluster sampling, and for two PSU per stratum designs.

Nascimento Silva and Skinner consider the problem of variable selection for regression estimation. They develop an approach based on minimizing the mean squared error of the resultant estimator. They empirically compare their approach to others using data from a 1988 test of Brazilian census procedures; the proposed procedures have good bias and mean squared error properties.

Eltinge and Yansaneh study the problem of formation of nonresponse adjustment cells. Within the general paradigms of estimated-probability and estimated-item based cells, they consider a variety of diagnostics for evaluating a set of adjustment cells. The diagnostic procedures include: comparison of estimates and standard errors for different numbers of adjustment cells; assessment of within-cell bias; assessment of cell widths relative to precision of estimated response probabilities; and comparisons of cell-based estimates to the unadjusted estimate.

Kovačević and Yung conduct an empirical study to compare variance estimation methods for measures of income inequality estimated from complex survey data. Variance estimation methods included in the study are: jackknife; bootstrap; grouped balanced half-sample method; repeatedly grouped balanced half-sample method; and a Taylor method based on estimating equations. After comparing relative bias, relative stability, and coverage properties of associated confidence intervals for a number of income inequality measures, they conclude that the Taylor method works best with the bootstrap method coming second.

Humphreys and Skinner investigate the use of the instrumental variable estimation method for estimation of gross flows among discrete states. This approach may be useful when external estimates of misclassification rates are not available. They numerically illustrate their method using data from the U.S. Panel Study of Income Dynamics and the two states "employed" and "not employed". They show that when measurement error is present, the unadjusted estimates can have considerable bias; this problem may be overcome by using suitable instrumental variables.

Waksberg, Judkins and Massey discuss issues involved in oversampling geographical areas to produce estimates for small domains of the population in demographic surveys, in conjunction with household screening. An empirical evaluation of the variance reduction is presented, along with an assessment of the sampling robustness over time. Simultaneous geographic oversampling for estimation of several small domains is discussed.

Losinger, in his paper, proposes a modified random groups standard error estimator for data from the U.S. Decenial Census sample. The usual random groups estimator has two undesirable properties for binomial variables: estimates of standard error for the "yes" and "no" responses are not equal; if all respondents answer "yes" the estimated standard error is not equal to zero. The essential idea of the proposed modification is to apply a ratio adjustment to each subgroup estimate so that subgroup estimates of population agree with the total.

Finally, Zeelenberg gives a simple technique, which exploits the use of differentials, to linearize design-based, nonlinear estimators. Ultimately, the linearized expressions allow one to obtain simple Taylor-based expressions for the variances of the nonlinear estimators. He illustrates the technique using two examples: the regression coefficient estimator and the regression estimator.

The Editor

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 3-10
Statistics Canada

A Computer Algebra for Sample Survey Theory

J.E. STAFFORD and D.R. BELLHOUSE[1]

ABSTRACT

A system of procedures that can be used to automate complicated algebraic calculations frequently encountered in sample survey theory is introduced. It is shown that three basic techniques in sampling theory depend on the repeated application of rules that give rise to partitions: the computation of expected values under any unistage sampling design, the determination of unbiased or consistent estimators under these designs and the calculation of Taylor series expansions. The methodology is illustrated here through applications to moment calculations of the sample mean, the ratio estimator and the regression estimator under the special case of simple random sampling without replacement. The innovation presented here is that calculations can now be performed instantaneously on a computer without error and without reliance on existing formulae which may be long and involved. One other immediate benefit of this is that calculations can be performed where no formulae presently exist. The computer code developed to implement this methodology is available via anonymous ftp at *fisher.stats.uwo.ca.*

KEY WORDS: *k*-statistics; Partitions; Product moments; Ratio and regression estimators; Symbolic computation; Variance estimation.

1. INTRODUCTION

In classical sampling theory two general problems concern us. These are the determination of an unbiased estimator of a parameter θ and the calculation of moments of $\hat{\theta}$, the estimator of θ.

The basic method to handle expectations and unbiased estimation is to operate on sample and population nested sums respectively through the inclusion probabilities, either single or joint probabilities as appropriate. A nested sum is a sum over the range of one or more indices such that each term in the sum depends on indices of different value. An unbiased estimator of any population nested sum is the associated sample nested sum with the quantity under the summation divided by the appropriate inclusion probability. Similarly the expectation of any sample nested sum is the associated population nested sum with the quantity under the summation multiplied by the appropriate inclusion probability.

In sampling theory, as well as several other areas of statistics, many algebraic calculations depend on a partition of some kind. With particular reference to sampling, Wishart (1952) showed that basic moment calculations under simple random sampling without replacement relied heavily on partitions. Here we will use partitions to express the sum of products of means or totals as linear combinations of nested sums and vice versa.

In the results presented here we consider the situation in which θ and $\hat{\theta}$ can be expressed as smooth functions of means or totals, population or sample as appropriate. There are two possibilities: the smooth function under consideration can be expressed as the sum of products of means or totals, or the smooth function cannot be so expressed. When the second possibility is operative the function $\hat{\theta}$ is first linearized through a Taylor expansion and θ is expressed as the root of an estimating equation. We use integer partitions to obtain terms in the Taylor linearization of a function or for the root of a function. The end result is that θ and $\hat{\theta}$ can be expressed, either exactly or approximately, as the sum of products of means or totals. These in turn can be expressed in terms of linear combinations of nested sums and vice versa. Estimation of θ or calculation of the moments of $\hat{\theta}$ is then a three step procedure: (a) Express an estimating equation for θ or the estimator $\hat{\theta}$ as the sum of products of means or totals, using Taylor linearization when necessary. (b) Transform the expression obtained in the first step to a linear combination of nested sums. Then operate on these nested sums to obtain unbiased estimates or expectations as appropriate. (c) Transform the resulting nested sums in the second step back into a sum a products of means or totals.

The key to automation of sampling theory results is the use of partitions. In general, whether these partitions are simple partitions, like that of an integer, or more complicated, like a full partition, each results from the repeated application of a fundamental rule. When the rule is identified, the possibility of automating a calculation arises. Seemingly unrelated formulae can result from the same fundamental rule and one computer algebra tool can be constructive in implementing many different calculations.

The notation used in the paper is outlined in §2. A discussion of expectation operators is given in §3. The concept of partitioning is reviewed in §4 and a rule is provided which leads to a simple recursive method for the enumeration of partitions. Integer partitions and Taylor linearization is discussed in §5. It is shown in §6 how the enumeration of partitions leads to the automatic calculation of expected values of products of sample means and *k*-statistics

[1] J.E. Stafford and D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

and to the derivation of unbiased estimators of products of finite populations means and k-statistics. Also in this section we apply the methodology to ratio and regression estimation.

Automation of these calculations and derivations will provide procedures which can be performed instantaneously and without error on a computer. Also, the reliance on formulae which may be long and involved is eliminated. A great deal of hand written algebra can be avoided. All computer code for the implementation of the methodology described here was written in the symbolic package *Mathematica* 2.0 which was installed on an IBM Risc 6000 with 64 megabytes of RAM. It is available via anonymous ftp at *fisher.stats.uwo.ca*. Although we use *Mathematica*, implementation in other environments such as *Maple*, *Macsyma* or *Reduce* is no doubt possible. For example, Kendall (1993) describes a system, implemented in *Reduce*, for the identification of invariant expressions. For a complete review of computer algebra in probability and statistics prior to 1991, see Kendall (1993).

2. SOME NOTATION

Consider a finite population of size N. A measurement of interest y_j is made on each unit $j, j \in U = \{1, ..., N\}$. In addition a single auxiliary variable x_j or possibly a $P \times 1$ vector of auxiliary variables $\boldsymbol{x}_j$ may be taken on the units. The p-th entry of this vector $\boldsymbol{x}_j$ is x_{pj}, where $p = 1, ..., P$. Several kinds of finite population parameters may be defined on the measurements y_j, x_j, or $\boldsymbol{x}_j$ for $j = 1, ..., N$. We denote a finite population parameter of interest by θ. Often θ can be expressed as a smooth function of finite population means, central moments and k-statistics. For convenience here we will deal only with means and k-statistics. Note that finite population variances and covariances are also second order k-statistics.

Not all N population elements are observed. Suppose that a sample s of size n is chosen from the population U by some sampling scheme. An estimator of θ, given by $\hat{\theta}$, is a smooth function of sample means and sample k-statistics.

In order to avoid much cumbersome summation notation we adapt the index notation of McCullagh (1987) to our purposes. For any j the vector $\boldsymbol{x}_j$ contain P entries so that each of these x-variables may be associated with one of the P indices. Suppose $\{i_1, ..., i_m\}$ is a subset of m of these P indices. In our adaptation of McCullagh's notation, $x_{i_1 j}$ is now what we called the vector $\boldsymbol{x}_j$. Products of these indexed quantities become multidimensional arrays. For example the product $x_{i_1 j} x_{i_2 j} x_{i_3 j}$ is a three-dimensional array of dimension $P \times P \times P$.

Let M denote a finite population mean. The argument of M shows the structure of the summand in the mean. For example, $M(y) = \sum_{j \in U} y_j / N$ and $M(yy)$ or equivalently $M(y^2) = \sum_{j \in U} y_j^2 / N$. In index notation, for example,

$$M(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in U} x_{i_1 j} x_{i_2 j} x_{i_3 j} / N \tag{1}$$

is a three-dimensional array. An element of this array is the mean of products in one of the permutations of the P elements taken three at a time in $\boldsymbol{x}_j$ where up to three of the elements may be alike. The (p,q,r)-th element of this array is $\sum_{j \in U} x_{pj} x_{qj} x_{rj}$ where $p, q, r = 1, ..., P$. The sample mean is denoted by m so that, for example,

$$m(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} / n. \tag{2}$$

For the purpose of making asymptotic expansions, since the variance of a given estimator $\hat{\theta}$ will be $O(n^{-1})$, we define a standardized variable for $\hat{\theta}$: it is the original variable $\hat{\theta}$ centered about its expectation and scaled by $1/\sqrt{n}$. That is,

$$z(\hat{\theta}) = [\hat{\theta} - E(\hat{\theta})] \sqrt{n}. \tag{3}$$

When necessary we use the summation convention of McCullagh (1987), where subscripts repeated as superscripts indicate implicit sums over that index. As a particular example, on assuming that the $\boldsymbol{x}_j$ are independent and identically distributed vectors from some infinite superpopulation, multivariate superpopulation moments can be obtained through the moment generating function which is expressed in this convention as

$$\mathrm{MGF}(\boldsymbol{t}) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h!, \tag{4}$$

where

$$\mu_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \ldots \partial t_{i_h}} \mathrm{MGF}(\boldsymbol{t})|_{\boldsymbol{t} = 0}. \tag{5}$$

By definition, the relationship between the moment generating function and the cumulant generating function is determined by the rule $\mathrm{MGF}(\boldsymbol{t}) = \exp\{\mathrm{K}(\boldsymbol{t})\}$, where

$$\mathrm{K}(\boldsymbol{t}) = \sum_{h=1}^{\infty} \kappa_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h! \tag{6}$$

is the cumulant generating function, where

$$\kappa_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \ldots \partial t_{i_h}} \mathrm{K}(\boldsymbol{t})|_{\boldsymbol{t} = 0}.$$

The finite population k-statistics, denoted by $K(\cdot)$, are defined as the unbiased (under the i.i.d. superpopulation model) estimators of the associated model cumulants. The number of arguments in K separated by commas denotes the order of the k-statistic. For example, the third order k-statistic $K(x_{i_1}, x_{i_2}, x_{i_3})$ is the model-unbiased estimate of (6), where

$$K(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{N}{(N-1)(N-2)} \times \sum_{j \in U} [x_{i_1 j} - M(x_{i_1})][x_{i_2 j} - M(x_{i_2})][x_{i_3 j} - M(x_{i_3})]. \tag{7}$$

In the univariate case finite population k-statistics are described in Wishart (1952). In particular $K(y,y)$ and $K(y,y,y)$ in the current notation are K_2 and K_3 in Wishart's (1952) notation. The sample k-statistics, denoted by $k(\cdot)$ with the appropriate arguments, are defined as the unbiased

estimators under simple random sampling without replacement of the associated finite population k-statistics. As in Wishart (1952) the sample k-statistic can be obtained from the population k-statistic upon replacing N by n and upon taking the sum over $j \in s$ rather than all units in the finite population. For example,

$$k(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{n}{(n-1)(n-2)} \times \sum_{j\in s} [x_{i_1 j} - m(x_{i_1})][x_{i_2 j} - m(x_{i_2})][x_{i_3 j} - m(x_{i_3})].$$

Note that if a comma is not present in the population or sample k-statistic, then the product of elements which appear together is required. For example, $K(xy)$ is the first order finite population k-statistic of a new variable which is the product of the measurements x_j and y_j for $j = 1,...,N$; $K(x,y)$ is a second order k-statistic, in particular the finite population covariance between x and y.

3. OPERATORS

The expectation operator E can be applied directly to any sample nested sum to obtain a finite population nested sum. Likewise an unbiased estimator of any finite population nested sum is a sample nested sum. In terms of triple nested sums, for example,

$$E\left[\sum_{J_3\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}\right] = \sum_{J_3=1}^{N} \pi_{jkl} x_{i_1 j} x_{i_2 k} x_{i_3 l} \tag{8}$$

and

$$\sum_{J_3=1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} \sim \sum_{J_3\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}/\pi_{jkl}, \tag{9}$$

where J_3 is the index set $\{j, k, l\}$ such that $j \neq k \neq l$ and where π_{jkl} is a joint inclusion probability. Parallel expressions may be established for with replacement sampling schemes.

Note that m will be unbiased for the associated M under simple random sampling without replacement. In general for any sampling design of fixed size n,

$$E[m(x_{i_1} x_{i_2} x_{i_3})] = \frac{N}{n} M(x_{i_1} x_{i_2} x_{i_3} \pi)$$

and

$$M(x_{i_1} x_{i_2} x_{i_3}) \sim \frac{n}{N} m(x_{i_1} x_{i_2} x_{i_3}/\pi)$$

where $M(x_{i_1} x_{i_2} x_{i_3})$ and $m(x_{i_1} x_{i_2} x_{i_3})$ are defined in (1) and (2) respectively.

The whole operation of finding expectation of an estimator $\hat{\theta}$ or of finding an unbiased estimator for the parameter of θ may be represented schematically as

$$\sum\prod \to \sum\sum \to \sum\prod, \tag{10}$$

where $\sum\prod$ denotes the sum of products and $\sum\sum$ denotes a sum of nested sums. If θ or $\hat{\theta}$ can be expressed as a $\sum\prod$ quantity, *i.e.*, a sum of products of means, then finding an unbiased estimator of θ or moments of $\hat{\theta}$ reduces to following the schema in (10) and applying the appropriate operator, such as those given in (8) or (9), to $\sum\sum$, the middle step in the schema. If θ or $\hat{\theta}$ are smooth functions of means but cannot be expressed directly as $\sum\prod$ quantities, then an initial step is required before applying the schema in (10). For $\hat{\theta}$ the initial step is to obtain a Taylor expansion of $\hat{\theta}$. For θ the initial step is to obtain an estimating equation and then to solve it for the parameter.

We illustrate the schema in (10) by considering the simple case of finding $E[\{m(x_{i_1})\}^2]$ under simple random sampling without replacement. The first operation is to express $\{m(x_{i_1})\}^2$ in terms of nested sums. In particular,

$$\{m(x_{i_1})\}^2 = \frac{1}{n^2}\sum_{j\in s} x_{i_1 j}^2 + \frac{1}{n^2}\sum_{j\neq k\in s} x_{i_1 j} x_{i_1 k}. \tag{11}$$

This is the $\sum\prod \to \sum\sum$ step. Now the expectation operator can be applied to $\sum\sum$. On applying inclusion probabilities $\pi_j = n/N$ and $\pi_{jk} = n(n-1)/[N(N-1)]$, the expectation operation on (11) yields

$$\frac{1}{n^2}\frac{n}{N}\sum_{j=1}^{N} x_{i_1 j}^2 + \frac{1}{n^2}\frac{n(n-1)}{N(N-1)}\sum_{j\neq k=1}^{N} x_{i_1 j} x_{i_1 k}. \tag{12}$$

Now the $\sum\sum \to \sum\prod$ step is applied. On expressing the nested sum in (12) as the sum of products, in particular $\sum_{j\neq k=1}^{N} x_{i_1 j} x_{i_1 k} = \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_1 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_1 j}$, the third operation yields

$$E[\{(m(x_{i_1})\}^2] = \frac{N(n-1)}{(N-1)n}\{M(x_{i_1})\}^2 + \frac{N-n}{n(N-1)} M(x_{i_1}^2). \tag{13}$$

In (13), $M(x_{i_1}) = K(x_{i_1})$ and $M(x_{i_1}^2) = [N/(N-1)]K(x_{i_1}, x_{i_1}) + K(x_{i_1})K(x_{i_1})$ so that (13) can be reexpressed as

$$E(m(x_{i_1})^2) = \{K(x_{i_1})\}^2 + (N-n)K(x_{i_1}, x_{i_1})/(Nn). \tag{14}$$

Likewise, following the schema in (10), the operations for finding an unbiased estimator of, for example, $\{M(x_{i_1})\}^2$ is similar to (11), (12) and (13). The estimand $\{M(x_{i_1})\}^2$ is expressed in nested sums similar to (11). These sums will be nested finite population sums. Similar to (12) the inclusion probabilities are applied. In this case the finite population sums are replaced by sample sums and summand is divided by the appropriate inclusion probability. Finally, similar to (13) the resulting nested sample sums are expressed as products of sums.

Each of the elementary operations to obtain an expected value through equations (11), (13) and (14), or to obtain an unbiased estimator, can be carried out using partitions. These operations are: expressing sums of products as nested sums and vice versa, and expressing means in terms of k-statistics and vice versa.

4. PARTITIONS AND FUNDAMENTAL PROCEDURES

Central to the automation of all algebraic calculations considered here is the notion of a partition. Partitioning as a focal point gives the appearance that the automated methods presented here are nothing more than an integer partition or a partition of an index set. While we assume that a partition of an integer is understood, a full partition requires a more formal definition.

Consider a set of m indices $I_m = \{i_1, ..., i_m\}$. A single partition P_m of I_m divides the m indices into $k \le m$ mutually exclusive and exhaustive subsets or blocks of I_m. We write $P_m = (b_1 \mid b_2 \mid ... \mid b_k)$, where the $b_1, ..., b_k$ are the blocks of I_m. P_m is unique up to permutations of indices within the blocks b_i. The block b_i is comprised of a subset of the indices of I_m. Elements within a block may be constrained to an alphabetical ordering and the blocks themselves may be ordered such that leading elements of each block are ordered alphabetically. This ensures the uniqueness of the partition P_m. In this case P_m would be called a standard ordered partition. Ordering the partitions in this manner does not offer any computational advantage and hence is not a requirement in what follows. The full partition of I_m is the set $\wp_m$ of all single partitions P_m of I_m.

Now we may identify the full partition of I_m in an algorithmic way via an inclusion-exclusion rule.

i. Let $\wp_1 = \{i_1\}$.
ii. An inclusion-exclusion rule determines the contribution to $\wp_t$ by a partition $P_{t-1} \in \wp_{t-1}$. In the inclusion part of the rule, the new index i_t is added as an element in turn to each of the blocks $b_1, ..., b_k$ which comprise P_{t-1}. If P_{t-1} has k blocks, k partitions for $\wp_t$ are created. In the exclusion part of the rule a new block containing the single index i_t is added to P_{t-1}.

For example, the full partition of $I_3 = \{i_1, i_2, i_3\}$ is given by the steps

i. $\wp_1 = \{(i_1)\}$
ii. $\wp_2 = \{(i_1 i_2), (i_1 | i_2)\}$ (15)
iii. $\wp_3 = \{(i_1 i_2 i_3), (i_1 i_2 \mid i_3), (i_1 i_3 \mid i_2), (i_1 \mid i_2 i_3), (i_1 \mid i_2 \mid i_3)\}$.

From step (i) to step (ii) the inclusion rule results in the partition $(i_1 i_2)$ and the exclusion rule results in $(i_1 \mid i_2)$. From step (ii) to step (iii) the inclusion rule results in the creation of the partitions $(i_1 i_2 i_3)$, $(i_1 i_3 \mid i_2)$, and $(i_1 \mid i_2 i_3)$. The exclusion rule yields the partitions $(i_1 i_2 \mid i_3)$ and $(i_1 \mid i_2 \mid i_3)$. This type of construction is easy to automate since it depends on a simple rule. Details of automating the partition of indices into full partitions and complementary set partitions are given in Stafford (1996).

Consider, for example, the classical problem of writing the model moments of the random vector x_{i_1} in terms of its cumulants. As in (5) we can identify the h-th moment array by differentiating MGF(t) in (4) h times and setting t equal to the zero vector. The result is the h-th coefficient in the expansion of MGF(t). Equivalently we can apply the same operation to $\exp\{K(t)\}$. In this case the result is a sum that depends on the coefficients of $K(t)$ in (6). For example, we may write the first three moments in terms of cumulants as follows:

$$\begin{aligned}
\mu_{i_1} &= \kappa_{i_1}\\
\mu_{i_1 i_2} &= \kappa_{i_1 i_2} + \kappa_{i_1}\kappa_{i_2}\\
\mu_{i_1 i_2 i_3} &= \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2}\kappa_{i_3} + \kappa_{i_1 i_3}\kappa_{i_2} + \kappa_{i_1}\kappa_{i_2 i_3} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3}.
\end{aligned}$$

Now in each case the result is a sum over the full partitions given in (15). These partitions arise since the multiplication rule for differentiation mimics the inclusion-exclusion rule for the enumeration of the full partition.

The above result is applied to sampling theory where we consider the problem of finding the expected value of a product of sample sums. The calculation requires expanding the product of the sums to identify terms where the finite population expectation operator will behave differently due to differences in the values of inclusion probabilities and joint inclusion probabilities.

For example, the product of sums $\sum_{j\in s} x_{i_1 j} \sum_{j \in s} x_{i_2 j} \sum_{j\in s} x_{i_3 j}$ can be expressed as

$$\sum_{j\in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 j} x_{i_3 k} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 k} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 k} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\neq l\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}. \quad (16)$$

The result corresponds to the full partition of the indices $I_3 = \{i_1, i_2, i_3\}$ given by $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order given for the terms in (16). For each partition in $\wp_3$, the variables in the same block have the same second index in the appropriate term in (16). For example, the partition $(i_1 i_3 \mid i_2)$ corresponds to the term $\sum_{j\neq k \in s} x_{i_1 j} x_{i_2 k} x_{i_3 j}$ in (16). Each term in the result can be identified by a partition of I_3 and each partition determines the manner in which the expected value operator will behave.

In general, we want to expand products of the form $\prod_{r=1}^{m} \sum_{j\in s} x_{i_r j}$, where the product is taken over the elements i_r of the index set $I_m = \{i_1, ..., i_m\}$. As in (16), the product can be expressed in terms of the full partition of I_m. This is because the iterative rule for expanding a product of sums mimics the inclusion-exclusion rule.

The expansion of the products of sums through partitions is demonstrated inductively as follows. Assume the product of the first $t-1$ sums can be expressed as a sum over the full partition of the index set $I_{t-1} = \{i_1, ..., i_{t-1}\}$, in particular

$$\prod_{r=1}^{t-1} \Big(\sum_{j\in s} x_{i_r j}\Big) = \sum_{P_{t-1}\in \wp_{t-1}} X_{P_{t-1}}. \quad (17)$$

In (17) the term $X_{P_{t-1}}$ is the sum identified by the partition $P_{t-1} = (b_1 | ... | b_k)$, $k = 1, ..., t-1$. The blocks b_j indicate groups of variables with the same second index and so P_{t-1} induces an index set $J_k = \{j_1, ..., j_k\}$ of second indices. We can express $X_{P_{t-1}}$ as

$$X_{P_{t-1}} = \sum_{j_1 \neq \ldots \neq j_k \in s} \left(\prod_{j \in J_k} X_{b_j} \right), \tag{18}$$

where X_{b_j} is a product of x's defined by the block b_j that all have the same second index. To illustrate (18), consider, for example, the third term of (16). Here $P_{t-1} = (i_1 i_3 | i_2)$ and $J_2 = \{j, k\}$ so that in (18) the sum is taken over $j \neq k \in s$ and the multiplicands of the product are $X_{b_j} = x_{i_1 j} x_{i_3 j}$ and $X_{b_k} = x_{i_2 k}$. Returning to the general discussion, when either side of (17) is multiplied by $\sum_{j \in s} x_{i_t j}$ the product of the first t sums is obtained. Now the product $X_{P_{t-1}} \sum_{j \in s} x_{i_t j}$ can be expressed as

$$\sum_{j_1 \neq \ldots \neq j_k \in s} \left(\sum_{l=1}^{k} x_{i_t j_l} \prod_{j \in J_k} X_{b_j} \right) + \sum_{j_1 \neq \ldots \neq j_k \neq j_{k+1} \in s} \left(\prod_{j \in J_k} X_{b_j} x_{i_t j_{k+1}} \right). \tag{19}$$

The first term in (19) corresponds to the inclusion part of the rule and the second term in (19) corresponds to the exclusion part of the rule. When (19) is summed over all $P_{t-1} \in \wp_{t-1}$, the result will be a sum over the full partition of the first t indices given by I_t, *i.e.*, the sum over all $P_t \in \wp_t$.

Once the product of sums, $\prod_{r=1}^{m} \sum_{j \in s} x_{i_r j}$, is expanded into a sum of nested sums, the finite population expected value operator can be applied to each term so that the expected value of this product can be obtained. The expected value under simple random sampling without replacement of the product of sums results in a weighted sum of nested sums, with each sum taken over the finite population. We then wish to evaluate these nested sums.

In general we wish to evaluate the nested sum $\sum_{J_t} Y_{J_t}$ where J_t is the index set $\{j_1, \ldots, j_t\}$. The sum is taken over all $j_1 \neq \ldots \neq j_t$ with each $j_r = 1, \ldots, N$. The summand Y_{J_t} is the product $x_{i_1 j_1} x_{i_2 j_2} \ldots x_{i_t j_t}$. In the special case when $t = 3$ or $J_3 = \{j, k, l\}$ the nested sum can be written in terms of full sums as

$$\sum_{J_3} Y_{jkl} = \sum_{j \neq k \neq l = 1}^{N} Y_{jkl} = \sum_{j \neq k \neq l = 1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} =$$

$$2\sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j} -$$

$$\sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} x_{i_3 j} + \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j}. \tag{20}$$

Note that the full sums in the rightmost expression in (20) result from the full partition $\wp_3$ in (15). The order of the partitions in $\wp_3$ is the same as the order of the terms on the right of (20). The subscripts on the right of (20) denote the block membership in $\wp_3$. For example, the partition $(i_1 i_3 | i_2)$ corresponds to the term $\sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j}$ in (20). Note also from (20) that the determination of a nested sum is complicated by the additional determination of the appropriate coefficients of the full sums.

In general the evaluation of finite population nested sums results from the repeated application of the rule

$$\sum_{j_1 \neq \ldots \neq j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{j_1 \neq \ldots \neq j_{t-1} = 1}^{N} \left[\prod_{r=1}^{t-1} x_{i_r j_r} \left(\sum_{j_t = 1}^{N} x_{i_t j_t} \right) \right]$$

$$- \sum_{j_1 \neq \ldots \neq j_{t-1} = 1}^{N} \left[\sum_{l=1}^{t-1} x_{i_t j_l} \left(\prod_{r=1}^{t-1} x_{i_r j_r} \right) \right]. \tag{21}$$

This expression mimics the inclusion-exclusion rule where the first set of sums on the right follows the exclusion part of the rule and the second set follows the inclusion part of the rule. Repeated application of (21) yields

$$\sum_{j_1 \neq \ldots \neq j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{P_t \in \wp_t} (-1)^{|J_t| - |P_t|}$$

$$\times \left\{ \prod_{b_k \in P_t} \left[(|b_k| - 1)! \sum_{j=1}^{N} \left(\prod_{i_r \in b_k} x_{i_r j} \right) \right] \right\}$$

where $|J_t|$, $|P_t|$ and $|b_k|$ are the number of indices in J_t, the number of blocks in the single partition P_t and the number of elements in the block b_k respectively.

5. INTEGER PARTITIONS AND TAYLOR LINEARIZATION

Suppose that under some sampling design an estimator $\hat{\theta}$ of a parameter θ is of interest. The methodology described in §§2 to 4 may be used in moment calculations for $\hat{\theta}$ or to find unbiased estimators of these moments. Only in the simplest cases can this methodology be applied directly. Typically $\hat{\theta}$ must be linearized so that it becomes a polynomial function of sample means or sums which are $O_p(1)$ random variables with respect to the sampling design. Once $\hat{\theta}$ is linearized in this way the methodology of §§2 to 4 is applicable.

The objective of the linearization is to write $\hat{\theta}$ as an asymptotic expansion where terms descend in order by $1/\sqrt{n}$, specifically

$$\hat{\theta} = \hat{\theta}_0 + \hat{\theta}_1/\sqrt{n} + \hat{\theta}_2/n + \ldots, \tag{22}$$

where $\hat{\theta}_i$ is the coefficient of the $n^{-i/2}$ term. Typically $\hat{\theta}$ is a product of quantities that can also be expanded in this way. For example, if the measurement of interest is y and one auxiliary variable x is present then θ might be $M(y)$ and the auxiliary information available is $M(x)$ as well as x_j for $j \in s$. Then $\hat{\theta} = M(x)m(y)/m(x)$, the simple ratio estimator, is a product of three quantities $M(x)$, $m(y)$ and $1/m(x)$ all having asymptotic expansions of their own. The expansion of $M(x)$ is itself. From (3) the expansion for $m(y)$ yields $M(y) + z(m(y))/\sqrt{n}$. The expansion for $1/m(x)$ results from (3) and then applying a Taylor expansion to $[M(x) + z(m(x))/\sqrt{n}]^{-1}$.

In general any expansion of a function with sufficient regularity can be found if operators are defined to expand a function, say $g(\dot{e})$ where $\dot{e}$ is itself an expansion. We are interested in expanding functions of the form

$$g(\dot{e}) = \prod_{j=1}^{p} \dot{e}_j. \tag{23}$$

where $\dot{e}_j$ itself has the expansion $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$. In linearizing $\hat{\theta}$ the basic requirement is to define an operator that returns $\hat{\theta}_i$ in (22). The efficiency of this operator derives solely from a rule for expanding functions of the form given in (23). The calculations required are functions of integer partitions. For example the $1/n$ term in the expansion of $\prod_{j=1}^{3} \dot{e}_j$ is

$$e_{21}e_{02}e_{03} + e_{01}e_{22}e_{03} + e_{01}e_{02}e_{23} + e_{11}e_{12}e_{13} + e_{11}e_{02}e_{13} + e_{01}e_{12}e_{13}. \tag{24}$$

Collecting first indices for each term in the sum results in a list in which each element sums to 2: {(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)}. On noting that the order $n^{-i/2}$ term in any expansion $\dot{e}_j$ is actually the (i + 1)-th term in the sum $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$, we may modify the list derived from (24) so that entries identify the position of terms in a sum. The modification is to add 1 to each index value in the list. In the list derived from (25) this results in all partitions of the integer 5 into 3 blocks: {(3,1,1), (1,3,1), (1,1,3), (2,2,1), (2,1,2), (1,2,2)}. In general, the i-th term in the expansion of $\prod_{j=1}^{p} \dot{e}_j$ or $\dot{e}_j^p$, where p is a positive integer, is a sum over all partitions of the integer $i + p$ into p blocks. Consequently, using this methodology any term in the expansion of, for example, the ratio estimator can be found.

We illustrate this technique with ratio and regression estimation. The ratio estimator is given by

$$M(x)\, m(y)/m(x) \tag{25}$$

and the regression estimator by

$$k(y) + b[K(x) - k(x)] = k(y) + \frac{k(x,y)}{k(x,x)}[K(x) - k(x)] \tag{26}$$

in the notation of k-statistics.

On using (3) the ratio estimator (25) may be expressed as

$$M(x)\left[M(y) + \frac{z(y)}{\sqrt{n}}\right]\left[M(x) + \frac{z(x)}{\sqrt{n}}\right]^{-1}. \tag{27}$$

The expression in (27) may be expressed in terms of (24) with $p = 3$. The first term in (27) is the expansion $\sum_{i=0}^{\infty} e_{i1} n^{-i/2}$ with $e_{01} = M(x)$ and $e_{11} = e_{21} = \cdots = 0$. The first term in square brackets in (28) is the expansion $\sum_{i=0}^{\infty} e_{i2} n^{-i/2}$ where $e_{02} = M(y)$, $e_{12} = z(m(y))$ and $e_{22} = e_{32} = \cdots = 0$. The second term in square brackets is the expansion $\sum_{i=0}^{\infty} e_{i3} n^{-i/2}$ where $\dot{e}_{i3} = (-1)^i \{z(m(y))\}^i / \{M(x)\}^{i+1}$. To get the $1/\sqrt{n}$ term in the expansion of (27), in which case $i = 1$ and $p = 3$, we need to find the integer partitions of 4 in blocks of 3. This yields the partitions (2,1,1), (1,2,1) and (1,1,2). On subtracting 1 from each index value in the list we obtain the list (1,0,0), (0,1,0), (0,0,1). Therefore the required term in the expansion is $(e_{11}e_{02}e_{03} + e_{01}e_{12}e_{03} + e_{01}e_{02}e_{13})/\sqrt{n}$ or equivalently $[z(m(y)) - M(y)z(m(x))/M(x)]/\sqrt{n}$. The $1/n$ term is obtained from (24) which reduces to

$$[M(y)\{z(x)\}^2/\{M(x)\}^2 - z(x)z(y)/M(x)]/n.$$

The regression estimator in (26) may be expressed as

$$K(y) + \frac{z(k(y))}{\sqrt{n}} + \left[K(x,y) + \frac{z(k(x,y))}{\sqrt{n}}\right] \times \left[K(x,x) + \frac{z(k(x,x))}{\sqrt{n}}\right]^{-1}\left[\frac{z(k(x))}{\sqrt{n}}\right] \tag{28}$$

using (3). The terms in the square brackets in (28) can be expanded in a similar fashion to the ratio estimator. In this case the terms in the expansions become: $e_{01} = K(x,y)$, $e_{11} = z(k(x,y))$ and $e_{21} = e_{31} = \cdots = 0$; $e_{i2} = (-1)^i\{z(k(x,x))\}^i / \{K(x,x)\}^{i+1}$ for $i = 0, 1, 2, \ldots$; and $e_{03} = 0$, $e_{13} = z(k(x))$ and $e_{23} = e_{33} = \cdots = 0$. Consequently, the $1/\sqrt{n}$ term in the expansion of the terms in the square brackets in (28) is

$$-\frac{K(x,y)z(k(x))}{K(x,x)\sqrt{n}}$$

and the $1/n$ term is

$$-\frac{1}{n}\left[\frac{z(k(x,y))}{K(x,x)} - \frac{K(x,y)z(k(x,x))}{K(x,x)^2}\right]z(k(x)).$$

These were obtained by the same argument that was used in the ratio estimator.

6. MACHINE APPLICATIONS TO THE CALCULATION OF EXPECTED VALUES OF SAMPLE STATISTICS AND THE DERIVATION OF UNBIASED ESTIMATORS

Since the machine application to the methodology described in §§3 to 5 was done in the programming language *Mathematica* we give a brief description of the operation of *Mathematica*. Then we describe the operators that were developed in *Mathematica* to provide a computer algebra for survey sampling theory.

Programming in *Mathematica* is carried out using expressions of the form $h[e_1, e_2, \ldots]$ where the object h is called the head of the expression and the e's are the elements of the expression. We have developed a number of machine expressions in *Mathematica* in the form of $h[e_1, e_2, \ldots]$ for operators which we apply to developing a computer algebra for sampling. All of these operators have been devised to

handle vectors as their arguments as well as scalars. There are four basic operators: $EV[\cdot]$ for expected value, $Cum[\cdot]$ for calculation of cumulants, $UE[\cdot]$ for unbiased estimator, and $Aexp[\cdot]$ for asymptotic expansion. There is also an operator to switch from notation using k-statistics to notation using means and vice versa.

The expected value operator $EV[\cdot]$ on sample statistics combines and carries out in *Mathematica* the three basic operations shown in the schema in (10). $EV[\cdot]$ contains two arguments, the first is the expression for which the expected value is to be obtained and the second is the sampling design which defines the inclusion probabilities. The application in *Mathematica* of $EV[\cdot]$ to $m(x_{i_1})m(x_{i_2})m(x_{i_3})$ under simple random sampling without replacement yields

$$K(x_{i_1})K(x_{i_2})K(x_{i_3}) + \frac{(N-n)(K(x_{i_1},x_{i_2})K(x_{i_3})}{Nn} + \frac{K(x_{i_1},x_{i_3})K(x_{i_2}) + K(x_{i_1})K(X_{i_2},x_{i_3}))}{Nn} + \frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2}$$

in the simplest expression of the output. Note that the result is a function of the full partition of $\{i_1,i_2,i_3\}$. If the operand is changed to $\{m(x_{i_1})-M(x_{i_1})\}\times\{m(x_{i_2})-M(x_{i_2})\}\times\{m(x_{i_3})-M(x_{i_3})\}$, application of $EV[\cdot]$ yields

$$\frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2},$$

which was obtained by Nath (1968) for particular values of the indices i_1, i_2 and i_3. In fact, the results in Nath (1968, 1969) for the products of three and four means and the exact results in Raghunandanan and Srinivasan (1973) for up to a product of eight means can all be reproduced automatically with the software that has been developed.

To this point the sampling design used in each of the examples has been simple random sampling without replacement. Results under general sampling designs can be obtained. We illustrate these results for the operator $Cum[\cdot]$ which is used to obtain the cumulants of an estimator. Note that the second cumulant for an estimator is also the variance. The operator $Cum[\cdot]$ has three arguments. The first is an expression for the estimator, the second is the order of the cumulant and the third is the sampling design. Under general sampling designs, estimators can be expressed in terms of $\sum\prod$ in the schema given by (10) and the $\sum\prod$ can be expanded to obtain $\sum\sum$, the middle term in (10). There is, however, no general simplification to obtain the final term in (10). This is illustrated with the Horvitz-Thompson estimator of $M(y)$ given by $(n/N)m(y/\pi)$ in the notation developed here. Application of the operator $Cum[\cdot]$ under a general sampling design to obtain the third cumulant of the Horvitz-Thompson estimator yields

$$2\frac{\left\{\sum_{i=1}^{N} y_i\right\}^3}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\frac{y_i^2}{\pi_i}\right\}}{N^3} - 3\frac{\sum_{i=1}^{N}\frac{y_i^3}{\pi_i^2}}{N^3}$$

$$-3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}y_iy_j}{(\pi_i\pi_j)}\right\}}{N^3} + 3\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}y_iy_j^2}{(\pi_i\pi_j^2)}}{N^3}$$

$$+\sum_{i=1}^{N}\sum_{h=1}^{N}\sum_{k=1}^{N}\frac{\dfrac{\pi_{ijk}y_iy_jy_k}{(\pi_i\pi_j\pi_k)}}{N^3}$$

where, for example, the term π_{ii} is the single inclusion probability π_i.

The operator $Aexp[\cdot]$ has two arguments, the function for which the expansion is required and the order of the expansion. This operator is used in combination with the $EV[\cdot]$ or $Cum[\cdot]$ operators to obtain approximate expectations or cumulants. This is illustrated in the case of the multiple linear regression estimator under simple random sampling without replacement. When there are q covariates the resulting regression estimator is given by

$$k(y) + b_{i_1}[K(x^{i_1}) - k(x^{i_1})] \tag{29}$$

using index and k-statistics notation. In (29) the coefficient b_{i_1} is the vector resulting from the product $k(x_{i_1}, y)\,ik(x^{i_1}, x_{i_2})$ in index notation, where the $q \times q$ array $ik(x_{i_1}, x_{i_2})$ is the inverse of the $q \times q$ array given by $k(x_{i_1}, x_{i_2})$. Similarly we will use $IK(x_{i_1}, x_{i_2})$ to denote the inverse of the finite population array $K(x_{i_1}, x_{i_2})$. Derivation of the mean square error of (29) requires Taylor expansions of the elements of b_{i_1} followed by the appropriate moment calculations and collection of terms. The *Mathematica* command to obtain the approximate variance of (29) is obtained by first applying $Aexp[\cdot]$ to (29) with 2 as the order in the expansion. Then the operator $Cum[\cdot]$ is applied to the result with the following arguments: the result from the asymptotic expansion as the estimator, simple random sampling as the design and 2 for the order of the cumulant. This yields

$$\frac{(N-n)K(y,y)}{Nn} + \frac{(-N+n)K(x_{i_1},y)K(x_{i_2},y)IK(x^{i_1},x^{i_2})}{Nn}$$

in index notation as output.

Estimation is achieved through the operator $UE[\cdot]$ which has two arguments, the estimand and the sampling design. For example, application of $UE[\cdot]$ to $\{M(x)\}^2$ under simple random sampling yields

$$\frac{(Nn)\{k(x)\}^2 + (N-n)k(x,x)}{Nn}.$$

If the estimand cannot be expressed as a sum of nested sums, but instead can be expressed as the root of an estimating function, then $UE[\cdot]$ obtains a consistent estimator.

7. DISCUSSION OF FUTURE WORK

The basic building blocks to develop a comprehensive computer algebra for survey sampling theory have been given. The foundation of this algebra is based on the enumeration of partitions. Fundamental operations under partition enumeration include the evaluation of nested sums and Taylor series expansions. Once these operations have been completed then expectations of sample statistics can be calculated or unbiased estimators of population quantities can be determined.

The next phase in this work is to extend the unistage results to multistage and multiphase sampling. In both multistage and multiphase sampling the problem reduces to the computer evaluation of multiple sums under an expectation operator or the determination of an unbiased estimator of multiple finite population sums. The problem of multistage sampling is currently under investigation. Another current area of inquiry is to extend the algebra to superpopulation models.

Once the basic algebra is in place then research problems involving algebraically complex sampling formulae can be easily investigated.

ACKNOWLEDGEMENTS

The authors are grateful to David Andrews for some useful discussions on this topic. This work was supported by grants from the Natural Sciences and Engineering Research Councils of Canada and by a research contract from Statistics Canada.

REFERENCES

ANDREWS, D.F., and STAFFORD, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society (B)*, 55, 613-628.

KENDALL, W.S. (1993). Computer algebra in probability and statistics. *Statistica Neerlandica*, 47, 9-25.

McCULLAGH, P. (1987). *Tensor Methods in Statistics*. New York: Chapman and Hall.

NATH, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

NATH, S.N. (1969). More results on product moments from a finite universe. *Journal of the American Statistical Association*, 64, 864-869.

RAGHUNANDANAN, K., and SRINIVASAN, R. (1973). Some product moments useful in sampling theory. *Journal of the American Statistical Association*, 68, 409-413.

STAFFORD, J.E. (1996). A note on symbolic Newton-Raphson, submitted for publication.

STAFFORD, J.E., and ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

WISHART, J. (1952). Moment coefficients of the *k*-statistics in samples from a finite population. *Biometrika*, 39, 1-13.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 11-21
Statistics Canada

Inverse Sampling Design Algorithms

SUSAN HINKINS, H. LOCK OH and FRITZ SCHEUREN[1]

ABSTRACT

In the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID). Important techniques, like regression and contingency table analysis, were developed largely in the IID world; hence, adjustments are needed to use them in complex survey settings. Rather than adjust the analysis, however, what is new in the present formulation is to draw a second sample from the original sample. In this second sample, the first set of selections are inverted, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so multiple simple random samples are drawn and a way to base inferences on them developed. Not all original samples can be inverted; but many practical special cases are discussed which cover a wide range of practices.

KEY WORDS: Finite population sampling; Inference in complex surveys; Resampling.

1. INTRODUCTION

The development of modern survey sampling is an extraordinary achievement (Bellhouse 1988; Hansen 1987; Kish 1995). The very richness in that development may have had the effect, though, of isolating survey sampling from the rest of statistics – where it is the richness of models that is given emphasis. In fact, it is a well-known commonplace that, in the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID).

Important techniques, like regression and contingency table analysis, were developed largely in this IID world; hence, adjustments are needed to use them in complex survey settings. Indeed, whole books have been written on this problem (Skinner, Holt and Smith 1989); and much time and effort have been devoted to it in software (like SUDAAN or WESVAR PC) specially written for surveys (See also Wolter 1985). With all that has been done already, can something more of value be added? We think we may have a contribution to offer on how to deal better with the "seam" which currently exists between IID and survey statistics.

Organizationally, the paper is divided into four sections. This introduction is Section 1. In Section 2 and 3 a general problem statement is provided and several "resolutions" are offered in a few of the better known designs. Our approach is to resample the complex sample to obtain an easier to analyze data structure. Specifically, we cover stratified element sampling, one and two-stage cluster samples, plus the important two PSU per stratum design (Section 2). Because any given resample is unlikely to contain all the information in the original survey, we look at what happens when the original complex sample is repeatedly resampled. A concrete illustration of our ideas is also given in Section 3; this has been taken from our practice and is based on a highly stratified Statistics of Income (SOI) sample of corporate tax returns (*e.g.*, Hughes, Mulrow, Hinkins, Collins and Uberall 1994). In a concluding section (Section 4), we discuss a few applications and some next steps needed for our still embryonic ideas to grow more useful.

2. PROBLEM STATEMENT AND POSSIBLE "RESOLUTIONS"

2.1 Motivation and Basic Approach

Suppose we wanted to apply an IID procedure to a complex survey sample. Suppose, too, that we wanted to take a fresh look at "solving" the seam problem that occurs because the survey design is not IID. How might one proceed? Well, there is a familiar expression that may fit our approach

If you only have a hammer, every problem turns into a nail.

Now, as samplers, we have a hammer and it is sampling itself. Can we turn the seam problem in surveys into a nail that can be dealt with by using another sampling design?

It is our contention that some of the time the answer to this question is "Yes." We call this second sample design an "Inverse Sampling Design Algorithm" – hence, the name of this paper.

A schematic might help visualize the algorithm (see figure 1). In the diagram two sampling approaches are compared – both yielding simple random samples from a population:

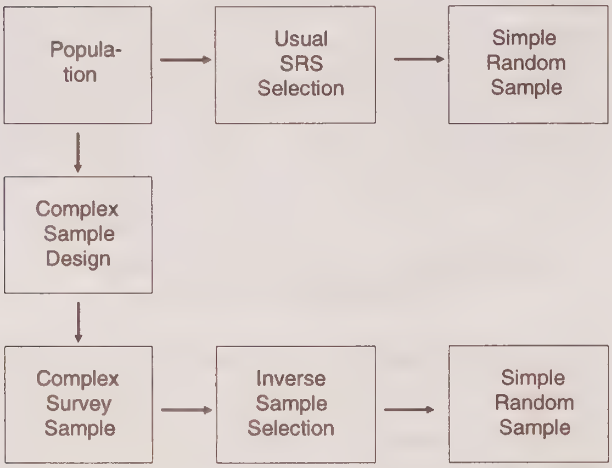
(1) The first design (top row) does this by employing a conventional direct simple random (SRS) selection process (*e.g.*, Cochrane 1977), such that all possible

[1] Susan Hinkins, Internal Revenue Service, Bozman, MT, U.S.A.; H. Lock Oh, Internal Revenue Service, Washington, DC, U.S.A.; Fritz Scheuren, Ernest and Young, 1402 Ruffner Rd., Alexandria, VA 22302 U.S.A.

samples of a given size have the same probability of selection. (Such designs are often impracticable or inefficient or both; hence, they are almost never used by survey samplers, despite their ubiquity in textbooks.)

(2) The second design envisions a two-step process. The first step is to sample the population in a complex way that focuses carefully on the nature of the population and the client's needs – using the client's resources frugally (this is the survey sampler's province, par excellence).

(3) What is new in our formulation is to draw a second (perhaps complex?) sample that inverts the first set of selections, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so we propose to create multiple simple random samples and base our inferences on them.



While elaborations are possible, the basic nature of the algorithms we are talking about should, by this point, be obvious. They can consist of just four basic steps:

(1) Invert, if you can, the existing complex design, so that simple random subsamples can be generated (to some useful degree of approximation).

(2) Potentially, apply your conventional statistical package directly to the subsample, since that is now appropriate.

(3) Repeat the subsampling and conventional analysis, in steps (1) and (2), over and over again.

(4) Retain, if you can, the flavour of the original randomization paradigm by using the distribution of subsample results as a basis of inference (rather than the original complex sample).

Notice some things that this approach is – and is not: First, it is extremely computer intensive – presupposing cheap, even very cheap computing. Second, it presupposes that practical inverse algorithms exist (which may not always be the case). Third, it also assumes that the original power of the full sample can be captured if enough subsamples are taken, so that no appreciable efficiency is lost. Fourth, as much as it may resemble the bootstrap (Efron 1979), we are not doing bootstrapping. There is no intent to mimic the original selections, as would be required to use the bootstrap properly (*e.g.*, McCarthy and Snowmen 1985; Rao and Wu 1988) – just the opposite; our goal here is to create a totally different and more analytically tractable set of subsamples from the original design.

2.2 Defining An Inverse Sampling Algorithm

Suppose that we wish to draw a simple random sample, without replacement, from a finite population of size N. Suppose further that the population is no longer available for sampling, but we have a sample selected from this population using a sample design D; let S_D denote this sample. Let S_m denote a second sample of size m that could be drawn from the population. An inverse sampling algorithm must describe how to select a sample from S_D so that for any given sample S_m

$$\Pr(\text{select } S_m \mid S_D) * \Pr(S_m \subset S_D) = \frac{1}{\binom{N}{m}}. \tag{1}$$

The first step is to calculate the probability that an arbitrary but fixed sample S_m is contained in the sample S_D. Obviously, there are constraints on the size of the simple random sample (SRS) that can be drawn in this manner; the probability that S_D contains S_m cannot be zero. Certainly, therefore, the SRS cannot be larger than the size of the original sample S_D, and in fact the size of the SRS is generally required to be much smaller than the original complex sample.

The problem, then, is to find a general algorithm to select an SRS from a given sample S_D with the correct conditional probability. It is also necessary to check that valid probability functions are used. The following subsections show the inverse sampling algorithms for a few of the more common sample designs: stratified, cluster, multistage, and stratified multistage designs. We also give an example where an inverse algorithm at first does not appear feasible.

2.3 Inverting A Stratified Sample

In this subsection the inverse algorithm is given for a stratified sample with four strata. The algorithm generalizes for any number of strata. We have a stratified sample with fixed sample sizes n_h in each stratum h, and known stratum population sizes, $N_1 + N_2 + N_3 + N_4 = N$. Because a given sample of arbitrary size m from the population might be contained entirely within one stratum, the largest simple random sample that can be selected from a stratified sample is of size $m = \min\{n_h\}$.

For a given sample S_m, let (x_1, x_2, x_3, x_4) denote the number of units in each stratum. Each x_i will be between 0 and m, and $x_1 + x_2 + x_3 + x_4 = m$. The probability that S_m is contained in the stratified sample is equal to the number of stratified samples containing these m units divided by the total number of possible stratified samples, *i.e.*

$$\Pr(S_m \subset S_D) = \frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}. \quad (2)$$

The algorithm for selecting a SRS from the stratified sample consists of the following three steps:

(1) Determine the size of the SRS to be selected: $m \le \min\{n_h\}$.

(2) Generate a realization $\{m_1, ..., m_4\}$ from a hypergeometric distribution, with probabilities

$$\Pr(m_1 = i_1, m_2 = i_2, ..., m_4 = i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}} \quad (3)$$

where $i_1 + i_2 + i_3 + i_4 = m$ and $0 \le i_1 \le m, 0 \le i_2 \le m, 0 \le i_3 \le m, 0 \le i_4 \le m$.

(3) In each stratum h, select a simple random sample of size m_h, without replacement, from the n_h sample units.

The conditional probability of selecting the sample S_m given that it is contained in the stratified sample, is then

$$\frac{\binom{N_1}{x_1}\cdots\binom{N_4}{x_4}}{\binom{N}{m}} \frac{1}{\binom{n_1}{x_1}\cdots\binom{n_4}{x_4}}. \quad (4)$$

The probability of selecting any given sample S_m using the inverse algorithm is the product of the two probabilities given in equations (2) and (4). It is straightforward to show that this product is equal to

$$\frac{1}{\binom{N}{m}}.$$

Therefore this procedure reproduces a simple random sampling mechanism unconditionally, *i.e.*, when taken over all possible stratified samples. Note that in order to generate all possible SRS's from this population, the entire sequence must be repeated, starting with selecting a stratified sample and proceeding through steps 1 - 3.

2.4 Inverting a One Stage Cluster Sample

In this subsection, we consider three special cases. To begin with, we examine cluster samples where the clusters are of equal size. This is followed by the more usual case where the clusters are of unequal size. In both of these settings we assume the clusters are sampled by a simple random sampling mechanism and without replacement. The third case studied is that of sampling unequal clusters by a probability proportional to size (PPS) mechanism. In this last instance we assume that the sampling is with replacement.

2.4.1 One Stage Cluster Sampling With Equal Cluster Sizes, Sampled With Equal Probability

Assume we have a population of N clusters where all clusters are of size M and k of them are selected by a simple random sampling mechanism without replacement.

To construct an inverse algorithm, we need to decide what the largest element subsample might be. It is immediate that the largest SRS of elements that can be selected is k. Incidentally, the cluster size is not a constraint on the size of the subsample.

For a given sample S_k, let q denote the number of clusters represented in S_k; $0 < q \le k$. Then the probability that S_k is contained in the cluster sample is equal to the number of cluster samples containing these q clusters divided by the total number of possible cluster samples, *i.e.*

$$\Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}. \quad (5)$$

As for the stratified sample, the algorithm first determines the number of units to be chosen from each cluster, $(m_1, m_2, ..., m_k)$. The probability distribution to be used to select the m_i's is

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1}\cdots\binom{M}{i_k}}{\binom{NM}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (6)$$

where $0 \le i_j \le k$, $i_1 + i_2 + ... + i_k = k$, and q is the number of nonzero i_j's. For example, with $M = 100$, $N = 6$, and $k = 3$

$$\Pr(m_1 = 1, m_2 = 0, m_3 = 2) = \frac{\binom{100}{1}\binom{100}{0}\binom{100}{2}}{\binom{600}{3}} * \frac{6*5}{3*2}$$

$$\Pr(m_1 = 3, m_2 = 0, m_3 = 0) = \frac{\binom{100}{3}}{\binom{600}{3}} * \frac{6}{3}.$$

Once the m_i's are determined, a simple random sample of size m_i is selected from cluster $i, i = 1, 2, ..., k$. Therefore the conditional probability of selecting S_k is

$$\Pr(\text{select } S_k \mid S_D) = \frac{1}{\binom{NM}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)}. \quad (7)$$

The probability of selecting a particular sample S_k is found by multiplying equation (5) times equation (7). It is routine to verify that this gives the correct probability of selecting an SRS.

Unlike the stratified example, where the function for selecting the values of m_i was a known probability function, it is not immediately obvious that equation (6) describes a probability distribution. Since the values generated by this function are all nonnegative, it need only be shown that they sum to one over the space of possible values. The first factor in the equation has the form of a hypergeometric distribution, except that the numerator is constrained to only k out of the N clusters, while the denominator still reflects the total N clusters. It is useful to define a partition of k as a combination of positive integers that adds to k, without regard to order. For example, the partitions of k = 3 are {3}, {1,2}, and {1,1,1}. Because the clusters are all of the same size, M, all patterns of selection that correspond to the same partition have the same probability of occurring. Take, for example, N = 6, and k = 3. In the full hypergeometric distribution, with equal cluster size, each of the following combinations has the same probability of occurring

$$(0,0,0,0,1,2), (0,0,0,0,2,1), (0,0,0,1,2,0), \ldots, (2,1,0,0,0,0).$$

The total number of such combinations is $N(N-1)\ldots(N-q+1)$, where q is the size of the partition, that is the number of (nonzero) values in the partition. In the example above, q = 2. For a given partition, if the nonzero counts can only be put into k specific cells, then there are $k(k-1)\ldots(k-q+1)$ such orderings. Therefore, summing the distribution over all values of $(i_1, \ldots, i_k)$ can be done by first summing over all partitions of k and then for each partition, summing over all possible orderings of that partition in k cells. Because all orderings associated with a particular partition share a common probability of occurrence, this results in a summation that is equivalent to summing the hypergeometric over the correct space, and therefore expression (6) sums to one.

The probability distribution needed for this simple cluster design (equation 6) is noticeably more difficult to generate than the hypergeometric distribution in the case of the stratified sample. However, as the sampling fraction k/N decreases, the probability is often contained in only two of the partitions: $q = k$ and $q = k - 1$. (These probabilities are calculated in the Appendix). Indeed, the probability may be concentrated in just the pattern with $q = k$ (A special case of this is also shown in the Appendix).

Given the results in the Appendix, it may be possible to approximate the exact inverse by selecting one case from each cluster, using systematic sampling from the original cluster sample. This approach is of real value because the probability distribution calculations become unwieldy as the number of clusters in the sample grows large. For a systematic inverse to work, however, the "step" would naturally have to be at least as large as M or maybe even greater, depending on the number of clusters in the population. To carry out this subsampling repeatedly, for each systematic sample inverse, the units within each cluster would be reordered randomly before the next selection and the clusters resorted randomly as well - then another random start obtained before stepping again through the original sample.

2.4.2 One Stage Cluster Sampling with Unequal Clusters, Sampled With Equal Probability

The inverse sampling algorithm for a sample of clusters of equal size does not generalize readily when a sample of unequal sized clusters is drawn. This is so despite the fact that it would appear to be straightforward to generalize this approach in an obvious way. In particular, it does not seem difficult to generalize the previous method so that the "probabilities" would multiply out successfully to give the "correct" probability of selection, *i.e.*

$$\frac{1}{\binom{M_+}{k}}, \quad \text{where } M_+ = \sum_1^N M_i. \quad (8)$$

However, generalizing to unequal cluster sizes M_i by selecting the m_i as

$$\Pr(m_1 = i_1, \ldots, m_k = i_k) = \frac{\binom{M_1}{i_1} \cdots \binom{M_k}{i_k}}{\binom{\sum_1^N M_i}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \quad (9)$$

does not result in a valid probability distribution. We will again assume, by the way, that the original clusters are being sampled with equal probability and without replacement, as was the case in subsection 2.4.1. Later (Subsection 2.4.3), as already noted, we will look at original samples which employ some form of Probability Proportional to Size (PPS) selection.

To see that it is not straightforward to simply generalize equation (6) into the form in equation (9), consider the following counter-example where the "probability" calculated using (9) is greater than one. Suppose N = 4 with cluster sizes; M_1 = 4, M_2 = 6, M_3 = 8, and M_4 = 10. Suppose further that we draw a cluster sample with k = 2 and that just by chance the two clusters picked are the largest – *i.e.*, M_3 = 8 and M_4 = 10. It is immediate that with these selections, equation (9) would generate a probability of selecting one unit from each cluster that was greater than one.

Can this difficulty be fixed? Yes, although not perhaps in an entirely satisfactory way. One method is to employ a

hypergeometric that assumes all the clusters were as large as the largest cluster in the population. The price paid is that the inverse sample size achieved is no longer fixed, and the resulting subsample is only conditionally SRS given the achieved sample size, denoted, say, as k_0. That is, for a given sample size k_0, $k_0 \le k$, all samples of size k_0 have the same probability of being selected using the inverse algorithm.

Let M_* denote the maximum cluster size, $M_* = \mathrm{Max}\{M_1, M_2, ..., M_N\}$. Create a population by filling out each original cluster with "dummy" units or placeholders, $j = M_i + 1, M_i + 2, ..., M_*$. Then using a method similar to Lahiri's (1951) for PPS sampling, the inverse algorithm selects units from the population consisting of N clusters each of size M_*, and then discards any element not in the "subpopulation" consisting of the original clusters of size M_i.

Specifically, given a cluster sample consisting of k clusters, select the vector $\boldsymbol{m}$ from the probability distribution

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M_*}{i_1}\binom{M_*}{i_2}\cdots\binom{M_*}{i_k}}{\binom{NM_*}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \tag{10}$$

where the components of $\boldsymbol{m}$ sum to k, and q of the components m_i are nonzero. This is now a proper probability distribution. Given the selected value of m_i, select a random sample of m_i units from cluster i, where the cluster contains M_i units from the population and $M_* - M_i$ "placeholders." Discard any selected units that are placeholders, in the set of $j = M_i + 1, M_i + 2, ..., M_*$. Therefore the final sample size will not necessarily be equal to k, but may be smaller, say k_0.

The resulting sample is conditionally a SRS from the population, in the sense that for a given value of k_0, all samples of size k_0 have the same probability of being selected using this inverse algorithm. To see this, continue to view the problem as a subpopulation, P, of N clusters of size M_i, $i = 1, ..., N$, within a population P_* of N clusters each of size M_*. Note that for any sample, S_*, of size k selected from the population P_*, the probability of selecting S_* using the inverse algorithm is

$$\frac{1}{\binom{NM_*}{k}}. \tag{11}$$

If $k_0 = k$ then this is the probability of selecting this sample using the inverse algorithm. For a fixed $k_0 < k$, let S_0 denote any given sample of size k_0 contained in P. We can generate a sample S_* containing S_0 by starting with S_0 and adding to it $k - k_0$ elements from the $N^* M_* - M_+$ placeholders in P_*. The number of such samples S_*, that result in selecting S_0, is

$$\binom{NM_* - M_+}{k - k_0} \quad \text{where} \quad M_+ = \sum_{i=1}^{N} M_i. \tag{12}$$

Therefore, the probability of selecting S_0 using the inverse algorithm is equal to the probability of selecting S_* using the inverse algorithm, given in (11), summed over all samples S_* constructed as described above, where the number of such samples is given by (12). This probability equals

$$\frac{\binom{NM_* - M_+}{k - k_0}}{\binom{NM_*}{k}}$$

and all samples of size k_0 have the same probability of being selected using the inverse algorithm.

There is a positive probability, unfortunately, that a sample might be selected with this approach that has no elements. This could occur if there were a large difference in the cluster sizes. However, if the number of clusters k in the original sample is large, this is unlikely to be a problem.

Again, as in the case of equal cluster sizes, an approximation is available using a systematic subsample as an inverse. This time we would want a step at least as large as the maximum cluster size. Using a systematic inverse, by the way, would have the advantage of controlling better the actual subsample size drawn.

2.4.3 One Stage Cluster Sampling with Unequal Clusters, Sampled With Unequal Probability

If a sample of k clusters is selected with PPS, an inverse algorithm may exist. Suppose the samples are selected with replacement from a population consisting of N clusters, with unequal cluster sizes, $M_1, M_2, ..., M_N$. Suppose, further, that the measure of size is either equal to M_i or proportional to M_i. Then at each draw,

$$\Pr(\text{select cluster } j) = \frac{M_j}{M_+} \quad \text{where } M_+ = \sum_{i=1}^{N} M_i. \tag{13}$$

Finally, since a one stage sample is being taken, once cluster j is selected, then all M_j units from that cluster are included in the sample.

An inverse algorithm in this case should result in a SRSWR. That is, for any vector $\boldsymbol{S}$ resulting from k independent selections from the population, the probability of selecting the ordered vector is

$$\Pr(\text{select } \boldsymbol{S}) = \left(\frac{1}{M_+}\right)^k. \tag{14}$$

An inverse algorithm is to simply randomly select one unit from each cluster in the cluster sample. Because the clusters were chosen with replacement, one should think of the sampled clusters as being ordered, by the order in which they were selected, or in any fixed order. For example, if the population contained 20 clusters, a possible cluster sample of size $k = 5$ is (7, 5, 7, 18, 6), *etc.*

The population consists of M_+ units, denoted as $u_1, u_2, \ldots, u_{M+}$. Let S denote a given sample, with replacement, $S = (s_1, s_2, \ldots, s_k)$, and let $\boldsymbol{c} = (c_1, c_2, \ldots, c_k)$ denote the associated cluster for each unit. For example, suppose the population is:

Cluster	Units
1	$u_1\ u_2\ u_3\ u_4$
2	$u_5\ u_6\ u_7\ u_8$
3	$u_9\ u_{10}\ u_{11}$
4	$u_{12}\ u_{13}\ u_{14}$
5	$u_{15}\ u_{16}\ u_{17}$
6	$u_{18}\ u_{19}\ u_{20}$

and $k = 3$. Then the sample $(s_1 = u_2, s_2 = u_4, s_3 = u_{17})$ corresponds to $\boldsymbol{c} = (1,1,5)$. The sample $(s_1 = u_{18}, s_2 = u_{19}, s_3 = u_{18})$ corresponds to $\boldsymbol{c} = (6, 6, 6)$. Note that this second sample can only be selected if cluster 6 is the only cluster chosen in the cluster sample.

For a given sample S of size k, and the corresponding vector $\boldsymbol{c}$ of cluster membership, the unconditional probability of selecting S using the inverse algorithm is

$$\Pr(\text{select } S \mid \text{cluster sample } \boldsymbol{c}) * \Pr(\text{select } \boldsymbol{c}) = \left(\prod_{i=1}^{k} \frac{1}{M_{c(i)}}\right)\left(\prod_{i=1}^{k} \frac{M_{c(i)}}{M_+}\right) \tag{15}$$

which is equal to the desired probability, equation (14).

Note that this same inverse algorithm works in the case where k clusters are selected with ppswr, but a sample of fixed size m is selected (srswor) from the chosen cluster, assuming that $M_i > m$ for all clusters i.

2.4.4 Some Comments On One Stage Designs.

We have seen that, with care, inverse algorithms can be constructed for several special cases where the original sample has a one stage cluster design. Two of our results are for cluster samples drawn with equal probability without replacement. The third is a ppswr design.

A convenient systematic inverse may even be workable as an approximation to the correct inverse algorithm when we have a cluster sample. The approximation works when using SRSWR is "close to" SRSWOR – *i.e.*, in our notation when k/NM is very small so that $1/(NM - k + 1)$ is approximately equal to $1/NM$. So everything seems intuitively to be consistent, across the cases studied.

Many cluster designs do not fall into any of the special cases examined. For some of them we conjecture that exact inverse algorithms may not exist. In particular, the general case of PPSWOR sampling seems to be one of these, including the frequently used variant of systematic PPSWOR. This may, or may not be a problem for practitioners who often employ the (usually) conservative practice of assuming that the sampling was with replacement – in which case an inverse algorithm would exist to the same order of approximation as was being assumed to estimate variances.

2.5 Multistage Cluster Designs

What about multistage designs? Can they be inverted? In some cases, we believe the answer is "Yes." Three designs will be looked at: (1) a two-stage design with simple random sampling at the first and second stages (Subsection 2.5.1); then, (2) a design which employed probability proportional to size (PPS) sampling at the first stage and simple random sampling at the second (Subsection 2.5.2). Finally, (3) the very important stratified multistage design with two PSUs per stratum deserves at least a brief comment.

As will be seen, the stratified and one stage results extend fairly readily. To demonstrate this, our basic strategy is to repeatedly apply the approaches already discussed earlier.

2.5.1 Multistage Designs With Simple Random Sampling at Both Stages

Suppose, first, that originally a simple random sample of k clusters, all of size M, was drawn at the first stage and a simple random subsample of size "r" was drawn at the second stage, within each cluster selected at the first stage.

As earlier, our inverse sample can be no larger than k. Suppose first that $1/(NM - k + 1)$ is approximately equal to $1/NM$, then we can employ an srswr inverse algorithm, since SRSWR and SRSWOR are very close. Using the results in Subsection 2.4.3, we would take a SRSWR sample of k clusters and then within each selected cluster take one observation at random. Alternatively, we could as in Subsection 2.4.1, first determine the number of units to be chosen from each cluster, $(m_1, m_2, \ldots, m_k)$. Once the m_i's are determined, a simple random sample without replacement of size m_i is selected from cluster i, $i = 1, 2, \ldots, k$. This may be a nearly exact result, except for the possibility that the inverse second stage sample size m_i may be larger than the original second stage sample size "r." When this occurs, we still can appeal to the results in Subsection 2.4.2 and draw our second stage sample with "placeholders." In this second instance, the resulting actual sample would no longer be fixed; but still would be conditionally SRS. If the first stage clusters are unequal in size but sampled with replacement, then we can again employ the trick used in Subsection 2.4.2 of creating "placeholders." The sample sizes are random and only conditionally do we achieve an SRS inverse.

Another way to approach this problem is to note that the largest SRS that can be selected using an inverse algorithm is

of size $k_0 = \min\{k, r\}$. This is done by first determining the number of units to select from each cluster, $(m_1, m_2, ..., m_k)$, where now the m_i's must sum to k_0 rather than k. Once the m_i's are determined, a simple random sample of size m_i is selected from cluster i, $i = 1, 2, ..., k$. The probability distribution to be used to select the m_i's is

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1} \cdots \binom{M}{i_k}}{\binom{NM}{k_0}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}$$

where $0 \le i_j \le k_0$, $i_1 + i_2 + ... + i_k = k_0$, and q is the number of nonzero i_j's.

One final comment, for both equal and unequal cluster sizes, the possibility of an approximate systematic inverse seems available – with essentially the same caveats, of course, as noted above.

2.5.2 Multistage Designs With PPS Sampling at the First Stage and SRS Sampling at the Second

Again, our inverse sample can be no larger than k. It is immediate that one way to construct an inverse would be to use the results in Subsection 2.4.3. Specifically, we would take a srswr sample of k clusters and then within each selected cluster take one observation at random. Other inverse algorithms may exist too. A systematic inverse seems reasonable, provided the probability of selecting the same cluster more than once is small to vanishing.

2.5.3 Stratified Multistage Designs With Two PSU's Per Stratum

Can two Primary Sampling Unit (PSU) designs be inverted? Our answer is "Yes," if the within stratum selections are made in one of the ways we discussed in detail earlier. This is basically the only case we will cover.

From our results in Subsections 2.3 and 2.4, it is immediate that if an inverse is to exist, then the sample size m cannot be any larger than $m = 2$. Depending on the sampling within each strata, we could employ one or more of the exact or approximate inverses to obtain two SRS selections within each stratum. To obtain an overall SRS sample, we would employ the inverse algorithm of Subsection 2.3 on these two selections and end up, finally, with just two selections overall.

2.5.4 Some Comments On Multistage Designs

In this Subsection, we have quickly covered a few multistage designs and provided exact or approximate inverses. The results were derived by appealing to earlier results in Subsections 2.3 and mainly 2.4. Of course, many multistage designs do not fall into any of the special cases examined - notably those with systematic selections at the last stage.

One last observation, many readers may wonder, at this point, how a method that selects only a sample of size two (as we did in Subsection 2.5.3) can be of any practical value. Perhaps the next section will help.

3. RESAMPLING TO INCREASE POWER

3.1 General Setting

Drawing a single, smaller simple random sample from a larger, more complex sample might be adequate for some users in some settings. However, for most users, the loss in power between the estimate based on the complex sample and the estimate based on a simple random sample would not be acceptable.

In order to increase the power of our approach, it was natural to consider resampling techniques. We are limited in the size of the SRS that can be drawn, but we can repeat the process. By repeating the entire subsampling procedure, we can generate g simple random samples each of size m, where each SRS is selected independently from the overall original sample. Each repetition must include all steps of the subsampling procedure. For example, in the stratified case, the stratum subsample sizes must be redrawn using the hypergeometric distribution.

In this section, conditions are given under which the precision of the estimates using multiple SRSs can be made arbitrarily close to the precision of the original estimates. We will begin our discussion by first defining some notation.

Let D denote any invertible design (such as a design of the type covered in Section 2). Let T be the population quantity of interest (say, a population total); and let T_D be an unbiased estimator of T calculated from the sample $\boldsymbol{S}_D$. Suppose g simple random samples are independently drawn from the given sample $\boldsymbol{S}_D$ and let t_i denote the estimator from the i-th simple random sample. Then it can be shown that

$$\text{if } E(t_i \mid S_D) = T_D$$

$$\text{then } \operatorname{Var}\left(\frac{1}{g}\sum_{i=1}^{g} t_i\right) = \operatorname{Var}(T_D) + \frac{1}{g}(\operatorname{Var}(t_1) - \operatorname{Var}(T_D)).$$

Proof: Because the g replications of the simple random sampling process are conditionally independent, then

$$\text{for } i \neq j, E(t_i t_j \mid \boldsymbol{S}_D) = T_D^{\,2}.$$

Therefore, unconditionally, for i not equal to j,

$$\begin{aligned}\operatorname{Cov}(t_i, t_j) &= E(t_i t_j) - T^2 \\ &= \operatorname{Var}(T_D).\end{aligned}$$

And the result follows directly.

Some of the conditions in this proof can be relaxed; if T_D is biased, then similar results can be obtained for MSE instead of variance. However, the condition that

$$E(t_i \mid S_D) = T_D$$

is necessary. And this condition is not met for ratio estimators. But, if the condition is met separately for the numerator and for the denominator of the ratio estimate and if the final size of the *combined* sample is sufficiently large so that a Taylor Series approximation is acceptable, then similar results can be found for approximations to the variance for ratios in the usual manner. Incidentally, even in the two PSU per stratum design, this approach works – provided we can obtain an unbiased estimate from each individual sample of size 2. And for estimates of totals, this can be the case – assuming at each stage of sampling that an inverse can be constructed.

3.2 Estimating The Sampling Error for Means or Totals

By resampling, one can achieve almost the same precision as the original design estimator. But because the resampled srs's are only conditionally independent, the estimation of the standard error is not as simple as if only one srs had been drawn. However the estimation remains relatively straightforward.

Let S^2 denote the population variance for the variable X and let T be its population total. For the sample means, totals and variances calculated from the generated simple random samples, let

$$t_{**} = \frac{1}{g}\sum_{j=1}^{g} t_j = \frac{1}{g}\sum_{j=1}^{g} N\bar{x}_j = \frac{1}{g}\sum_{j=1}^{g}\frac{N}{m}\sum_{i=1}^{m} x_{ji}$$

$$s_j^2 = \left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(x_{ji} - \bar{x}_j)^2$$

$$s_*^2 = \left(\frac{1}{gm-1}\right)\sum_{j=1}^{g}\sum_{i=1}^{m}(x_{ji} - \bar{x}_{**})^2$$

where $\bar{x}_{**} = \dfrac{t_{**}}{N} = \dfrac{1}{gm}\sum_j\sum_i x_{ji}$.

Note that the sample variance using all gm units can be expressed as

$$s_*^2 = \frac{1}{mg-1}\left[(m-1)\sum_{j=1}^{g} s_j^2 + \frac{m}{N^2}\sum_{j=1}^{g}(t_j - T)^2 - \frac{mg}{N^2}(t_{**} - T)^2\right].$$

Hence

$$E(s_*^2) = \frac{1}{mg-1}\left[g(m-1)S^2 + \frac{m}{N^2}\sum_{j=1}^{g}\mathrm{Var}(t_j) - \frac{mg}{N^2}\mathrm{Var}(t_{**})\right].$$

Rewriting this gives

$$\mathrm{Var}(t_{**}) = N^2\left(\frac{m-1}{m}\right)S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{mg-1}{mg}\right)E(s_*^2).$$

Therefore, by replacing S^2 and $\mathrm{Var}(t_i)$ with unbiased estimates and replacing $E(s_*^2)$ with s_*^2, we can generate approximately unbiased estimates of $\mathrm{Var}(t_{**})$.

It may be worth emphasizing that this result does not require the user to know anything about the original sample design. If users are given a way to invert the original design, then they can, by repeated subsampling, achieve nearly the efficiency of the original design and readily estimate the appropriate sampling errors. There is one condition on this result, namely that the subsample size be such that $m \geq 2$. Incidentally, for $m = 2$, the variance expression becomes

$$\mathrm{Var}(t_{**}) = \frac{N^2}{2}S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{2g-1}{2g}\right)E(s_*^2).$$

Based on this, as above, a variance estimator could be built for two PSU per stratum designs.

3.3 An SOI Illustration

In this subsection we consider an example of an inverse algorithm and how well it works. The Statistics of Income (SOI) corporate sample will be our starting point. Now, as noted earlier, the SOI sample has essentially a stratified SRS design and so can be inverted (subsection 2.2).

It is our belief that many SOI users might find a full SRS inverse sample more valuable and easier to employ than the complete, stratified sample data base. An interim goal could be to provide them with a set of simple random samples. A more flexible system would be to provide the interactive software to allow the user to designate the simple random samples of interest, to be selected from the complete data base.

In our simulations we used four of the strata in the SOI sample of corporate returns, namely the strata representing the smallest regular corporations (Hughes *et al.* 1994). As can be seen from table 1, the stratified sample (of four strata) consisted of 15,618 units, and the largest SRS that can be selected is $m = 2{,}224$. The table also shows the population sizes and the estimated variance of the variable Total Assets, within each stratum.

Table 1
Corporate Population and Sample Size, plus Estimated Stratum Variances, For Four SOI Stratum

Strata (h)	N_h	n_h	S_h^2 (in 1000's)
1	1,376,801	3,889	222,808
2	552,909	2,224	670,162
3	678,371	4,005	12,796,578
4	436,023	5,500	14,984,753

The variable total assets was used because it is the primary stratifying variable; and, therefore, the loss in precision due to removing the stratification should be relatively large. Indeed, this proved to be the case.

Shown below is the ratio of the variance of the estimated total using g simple random samples, of 2,224 each, divided by the variance of the total based on the stratified sample. The table displays values of g from 1 to 1,000. For example, if only one SRS is selected the variance of the estimated total is 29 times larger than the variance of the stratified total.

g	Relative Variance Increase
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

By resampling 500 to 1,000 times, the variance has been reduced to the same order of magnitude as the stratified sample. Even at 100 subsamples good results exist here, suggesting that the use of an inverse algorithm could work well for strata such as these. This is not to recommend that an inverse algorithm be employed in general with so few resamples. Doubtless, in highly skewed populations a much larger number would be required.

4. POTENTIAL APPLICATIONS AND NEXT STEPS

In this paper we have shown that inverse sample design algorithms exist in a few special cases. We do not, as yet, have a general result – if, indeed, there is one. This is clearly a part of the problem that needs more work. Like most tools, an inverse sampling algorithm may not be the best choice in certain cases; it may not be even a reasonable alternative in some circumstances. But there are applications where it appears to have advantages and so should be considered. In this section we both briefly suggest areas where this methodology may be useful and also mention some of the limitations and problems that remain.

Customer-Driven Perspective – It is worth emphasizing the customer-driven nature of our approach. Even if it could not be justified on other grounds, inverse algorithms might be advocated as a part of "reinvention" (*e.g.*, Osborne and Gaebler 1992). Right now many large complex surveys may not be sufficiently benefiting society, because they are so badly under-analyzed or even misanalyzed:

- For the long run, we must work towards increasing the survey and general quantitative literacy of existing and potential customers – *e.g.*, as with the new series *What Is a Survey?* (Scheuren (ed.) 1995).
- In the short run, we need to start where our customers are – giving due respect to the often small part that survey data may add to their decision making. Certainly it is worth thinking about ways to lower the cognitive costs customers bear when using our complex survey "products."

A "Sample" of Possible Opportunities – There is an increasing awareness of the weaknesses within the traditional randomization paradigm (*e.g.*, Särndal and Swensson 1993). Of particular concern here is all the fiddling we have to do when trying to correct for nonsampling errors. Some of this flavour is evident in Rao and Shao (1993). By putting the possible adjustments for these nonsampling errors back into a simple random sampling framework, we may, indeed, be able to make more progress.

For decades, survey practitioners have elaborated exceedingly complex sample designs; and, then, made efficient point and confidence interval estimates from them. On the other hand, how much do we really understand about the distributions that our sample estimators generate when effective sample sizes are small to moderate? Will we be able to fully capitalize on the "visualization revolution" now occurring (*e.g.*, Cleveland 1993)? Particularly in the presence of nonsampling error? Maybe we should be building in a way to always look at distributions. The use of an inverse sampling algorithm might be one possibility (See also Pfeffermann and Nathan 1985). In any case, stronger visualization tools for complex surveys could help, even the very experienced among us, deepen our intuitions and connect them better to the particular population under study. Obviously, visualization efforts also pay off by lowering the price customers pay to use survey data.

An intriguing problem where the inverse sampling algorithm may have an application is the case where we have a two PSU per stratum design with L strata where L is small, say less than 30. Suppose further that for some of the variables in the survey the stratification and clustering are unimportant – *i.e.*, the design effect is $\delta = 1$, approximately. For these variables, would it not be possible for the stability of the variance estimate to be greater with the resampled method than with the Balanced Repeated Replication (BRR) approach to variance estimation that is usually employed?

Another example that we are considering is the case where the user is interested in tests of independence in 2×2 tables, based on stratified sample data (Hinkins, Oh and Scheuren 1995). For the chi-square test statistic we are now in the midst of comparing our results with the approach suggested by Scheuren (1972) and Fellegi (1980). So far it appears that the power of our method is comparable to these more familiar approaches (as might be expected from, say, Westfall and Young (1993)). This may be an instance where the extra work involved in the inverse sampling algorithm may have real benefits – beyond just making it easier for users to employ familiar tools – by allowing the user to look at the distribution rather than just one p-value.

A "Sample" of Problems Remaining – A "sample" of the problems that remain with our inverse algorithm might be given here. For example, what happens when we do not know what the population size is? What happens when the population has more than one elementary unit – persons, say, for one analysis; households for another; neighbourhoods for still a third? Answers exist for these difficulties but they have

an *ad hoc* flavour to us. In many surveys, for instance, we guess about N and use that guess in poststratification. That degree of approximation for an inverse might be acceptable. For the problem of multiple analysis units, we could do several inverses. While potentially workable, this seems exceedingly awkward.

We have indicated that in some cases it may not be too difficult to resample multiple times using the inverse algorithm in order to reproduce reasonable efficiency. But what about the case where the user of a stratified sample is interested in subpopulations. If the domains of interest are in fact the strata, then the user does not gain any benefits by using the SRS's produced using the inverse algorithm. If the domains of interest cut across the strata and they are small, then the number of samples required using the inverse algorithm may be very large in order to maintain reasonable estimation for the domains.

Finally, we briefly mention one more problem that we have thought about. Many multistage designs actually select only one PSU per stratum. The strata are then paired for variance estimation purposes. We have already noted that an inverse to this approximation is available which can be made about as good as that approximation is to begin with. Is there a way to get a better approximation using the inverse approach directly?

Last Words – Many things are changing in our profession. The worldwide quality revolution certainly has had an impact (Mulrow and Scheuren 1996). We are remaking the way surveys are done – from design, to data capture, to the way customers use them. This paper may be a small contribution to that process.

ACKNOWLEDGEMENTS

We wish to express our particular appreciation to the referees and associate editor for their insightful prodding and scholarship. The original submission we sent in was only a sketch of what is now included. We also owe a debt of gratitude to Phil Kott, who has been discussing our ongoing work at various Washington Statistical Society meetings.

APPENDIX

Suppose one has a cluster sample of k clusters from a population of N clusters, where each cluster has the same number of units, M. In the inverse sampling algorithm, the first step is to choose the vector $(m_1, m_2, ..., m_k)$ containing the number of units to be chosen from each cluster. Let q indicate the number of nonzero values of m_i. The probability of selecting the one pattern with $q = k$, that is the pattern with $m_i = 1$, for all $i = 1, 2, ..., k$, is

$$\Pr(q = k) = M^{k-1} \frac{(N-1)(N-2)...(N-k+1)}{(NM-1)(NM-2)...(NM-k+1)}.$$

Call this probability P_1. If $NM >>> k$ then P_1 can be approximated by

$$\prod_{i=1}^{k-1} \frac{(N-i)}{N} = \frac{(N-1)(N-2)...(N-k+1)}{N^{k-1}}.$$

Consider next the partition of k corresponding to $q = k - 1$; this corresponds to exactly one partition of k, namely $\{1, 1, ..., 1, 2\}$. There are $k(k-1)$ equally likely possible patterns of $(m_1, ..., m_k)$ with $q = k - 1$. The probability of selecting a vector $\boldsymbol{m}$ with $q = k - 1$, is

$$\Pr(q = k - 1) = \frac{k(k-1)(M-1)}{2M(N-k+1)} P_1.$$

Therefore it is not difficult to calculate the probability that the selected $\boldsymbol{m}$ has either $q = k$ or $q = k - 1$. The following table shows some examples for two values of M.

Table A
$\Pr(q = k-1 \text{ or } q = k)$

k	N	$M = 10$	$M = 100$
4	8	.92	.90
4	20	.99	.98
10	20	.38	.34
10	30	.63	.59
10	50	.83	.80
10	200	.99	.98
50	500	.35	.30
50	1000	.70	.66
50	5000	.98	.98

For small k, it is not difficult to calculate the entire probability distribution needed to generate $\boldsymbol{m}$. But as k increases, the number of partitions increases, and this calculation becomes difficult or at least tedious. For $k = 4$, there are only 4 partitions; for $k = 10$ there are 39 possible partitions. One can see from Table A, that as the cluster sample becomes "larger," if the sampling rate is small enough, *i.e.*, if $k << N$, then one might only need to calculate the probabilities for these two partitions in order to approximately invert the cluster sample. For $k = 10$ and $N = 200$, these two partitions essentially account for all of the probability distribution.

The probability of selecting just one unit per cluster ($q = k$) is smaller than the values in Table A; so, in order to use a systematic inverse, we would want $k <<< N$. This can be obtained in some settings when the number of clusters is large and we are willing to take k very small, relying on repeatedly resampling the original survey, as described in Section 3.

To illustrate, assume a sample of size k_0 where, of course, $k_0 < k$, so that an inverse is possible; Further, to see if a systematic inverse would work, let $k_0 <<< N$. This is the case we illustrate in table B. In table B, we have confined

attention to just one value of N, N = 5000 clusters, although the results could be extended readily.

Table B

Pr{inverse sample picks the pattern (1,1, ..., 1)}

k_0	k_0/N	$M = 10$	$M = 100$
2	.0004	.9998	.9998
5	.001	.9982	.9980
10	.002	.9919	.9911
20	.004	.9663	.9627
30	.006	.9245	.9166
40	.008	.8687	.8553
50	.01	.8015	.7821

Clearly, as k/N gets small, a systematic sample becomes a better and better approximate inverse. Only experience would confirm if the approximation at $k_0 = 20$ and $k_0/N = .004$, say, is adequate. We think it might be, especially since the effect of using a systematic inverse usually is to make the variance calculations more conservative (since typically the intracluster correlation $\rho > 0$).

REFERENCES

BELLHOUSE, D. (1988). A brief history of random sampling methods. *Handbook of Statistics,* 6, 1-14.

CLEVELAND, W. (1993). *Visualizing Data.* Summit, NJ: Hobart Press.

COCHRAN, W. (1977). *Sampling Techniques.* New York: Wiley.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7, 139-172.*

FELLEGI, I. (1980). Approximate tests of independence and goodness of fit based on multistage samples. *Journal of the American Statistical Association,* 75, 261-268.

HANSEN, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 162-179.

HINKINS, S., OH, H.L., and SCHEUREN, F. (1995). Using an Inverse Algorithm for Testing of Independence Based on Stratified Samples. George Washington University Technical Report.

HUGHES, S., MULROW, J., HINKINS, S., COLLINS, R., and UBERALL, B. (1994). Section 3, *Statistics of Income – 1991, Corporation Income Tax Returns*, 9-17. Washington, DC: Internal Revenue Service.

KISH, L. (1995). The Hundred Years Wars of Survey Sampling. Centennial representative Sampling Conference, Rome, May 31, 1995.

LAHIRI, D. (1951). A method for sample selection providing unbiased ratio estimates, *Bulletin of the International Statistical. Institute*, 34, 72-86.

McCARTHY, P., and SNOWDEN, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics.* Series 2, No. 95, DHHS Pub. No. (PHS) 85-1369. Washington, DC: Public Health Service.

MULROW, J., and SCHEUREN, F. (1996). Measuring to improve quality and productivity in a processing environment. *Data Quality*, 2, 11-20.

OSBORNE, D., and GAEBLER, T. (1992). *Reinventing Government.* New York: Plume.

PFEFFERMANN, D., and NATHAN, G. (1985). Problems in model identification based on data from complex samples. *Bulletin of the International Statistical Institute*, 68.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

SÄRNDAL, C.-E., and SWENSSON, B. (1993). Washington Statistical Society talk on the shifting nature of the survey sampling paradigm.

SCHEUREN, F. (1972). Topics in Multivariate Finite Population Sampling and Data Analysis. George Washington University Doctoral Dissertation.

SCHEUREN, F. (Ed.) (1995). *What is a Survey?* One of a series of pamphlets published by the American Statistical Association to increase survey literacy.

SKINNER, C., HOLT, D., and SMITH, T., (Eds.) (1989). *Analysis of Complex Surveys.* New York: Wiley.

WESTFALL, P., and YOUNG, S. (1993). *Resampling-Based Multiple Testing*. New York: Wiley.

WOLTER, K. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 23-32
Statistics Canada

Variable Selection for Regression Estimation in Finite Populations

PEDRO L.D. NASCIMENTO SILVA and CHRIS J. SKINNER[1]

ABSTRACT

The selection of auxiliary variables is considered for regression estimation in finite populations under a simple random sampling design. This problem is a basic one for model-based and model-assisted survey sampling approaches and is of practical importance when the number of variables available is large. An approach is developed in which a mean squared error estimator is minimised. This approach is compared to alternative approaches using a fixed set of auxiliary variables, a conventional significance test criterion, a condition number reduction approach and a ridge regression approach. The proposed approach is found to perform well in terms of efficiency. It is noted that the variable selection approach affects the properties of standard variance estimators and thus leads to a problem of variance estimation.

KEY WORDS: Auxiliary information; Calibration; Sample surveys; Subset selection; Ridge regression.

1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977, chap. 7). For the basic case when the population mean $\bar{X}$ of a vector of variables x_i is known and simple random sampling is used, the regression estimator of the population mean $\bar{Y}$ of a survey variable y_i takes the form

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})' b \tag{1}$$

where $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i respectively, and b is the sample vector of linear regression coefficients of y_i on x_i.

Regression estimation is useful for at least three reasons. First, it is flexible. Any number of population means of continuous or binary variables can, in principle, be incorporated into $\bar{X}$. In particular, poststratification arises as a special case (Särndal, Swensson and Wretman 1992, sec. 7.6). The procedure also extends to handle complex sampling designs. Second, regression estimation has certain optimal efficiency properties. See, for example, Isaki and Fuller (1982, Theorem 3). Third, $\bar{y}_r$ has the "calibration" property that if y_i is one of the variables of x_i so that $\bar{Y}$ is known then $\bar{y}_r = \bar{Y}$ (Deville and Särndal 1992).

In this paper we consider the question of how to select the x variables for use in the regression estimator. This question is of interest for at least two reasons. First, there is simply the practical reason that in some circumstances the number of potential variables in x_i may be very large. For example, in population censuses in a number of countries values of some variables are recorded on a "short form" for all individuals and values of other variables are collected on a "long form" for a sample. The population means of the short form variables together with their squares, cubes, products and so forth will thus be known. Small area identification will also typically be available. Thus the dimension of x_i as a vector containing functions of the short form variables together with dummy variables representing each small area could easily run into the thousands. In such cases, the selection of x variables becomes a practical necessity.

A second reason is more fundamental for a model-assisted or model-based approach to survey sampling. These approaches may be characterised as follows in the context of regression estimation. First a regression model is selected which has "good predictive power", so that the regression estimator will have "good efficiency". Then, either a design-based approach to inference is adopted in the model-assisted approach (Särndal *et al.* 1992) or model-based prediction is employed in the model-based approach. Although the literature on the latter problem of inference is vast, there seems remarkably little formal attention devoted to the former model selection problem. In practice, the most that seems to happen is that the "main" x variables which account for "most of" the sample R^2 are chosen (*cf.* Särndal *et al.* 1992, sec. 7.9.1). However, more theoretical guidance seems needed, especially when a large number of x variables is available.

A further reason for considering the variable selection problem more formally is that it may help clarify the issue of the impact of variable selection on inference. The problem that sample-based selection of estimators may affect the properties of the selected estimator has long been recognized (Hansen and Tepping 1969, App.) but little study seems to have been made of what the effects may be.

In this paper we consider a variable selection approach aimed at minimising the mean squared error of $\bar{y}_r$. First, however, we study the dependence of the mean squared error of $\bar{y}_r$ on the number of x variables in section 2 and then consider alternative estimators of the mean squared error of $\bar{y}_r$

[1] Pedro L.D. Nascimento Silva, IBGE-Departamento de Metodologia, Avenida Chile 500, Rio de Janeiro-RJ, Brasil; and Professor Chris J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

in section 3. Variable selection procedures based on these estimators are then proposed in section 4.

We contrast our variable selection approach with four existing approaches. First, we consider the traditional approach of using a fixed subset of auxiliary variables regardless of the observed sample. Next, we consider a "condition number reduction procedure" inspired by work of Bankier (1990), in which auxiliary variables are discarded in order to reduce the condition number of a certain cross-products matrix of the x variables.

Third, we follow Bardsley and Chambers (1984) and consider a ridge regression approach. This does not involve variable selection but instead addresses the possible problem of multicollinearity in the regression estimator by modifying the estimator, allowing for some calibration error. Both the ridge regression and condition number reduction procedures have the advantage that they do not require specification of a response variable y, because they aim to provide a single set of "calibration" weights to be used for all survey variables. However, they do not guarantee gains in efficiency. Their results are separated by a line from the results for the other procedures in the tables presented in section 6 to indicate that they differ.

Fourth, we consider variable selection following conventional significance test criteria. Our general view is that the objective of variable selection in regression estimation for finite populations is quite different from the objective of parameter estimation or prediction of y values for single observations in classical regression (Miller 1990). However, it seems desirable to treat such an approach as one benchmark for comparison.

In section 5 we consider properties of the regression estimator following variable selection on the basis of estimated variances. Section 6 describes an empirical study carried out to compare our proposed variable selection procedures with the competing procedures described above. This study used data from a test census carried out in the municipality of Limeira, Brasil, as part of the preparation for the 1991 Brazilian Population Census. Section 7 presents our conclusions and some directions for further research.

2. THE DEPENDENCE OF THE VARIANCE OF THE REGRESSION ESTIMATOR ON THE NUMBER OF x VARIABLES

We begin by defining some notation. Let $U = \{1, \ldots, N\}$ denote a finite population of N distinguishable elements and let $s \subset U$ denote a sample of n distinct elements drawn from U according to a simple random sampling without replacement design. Let $x_i = (x_{i1}, \ldots, x_{iq})'$ be the $q \times 1$ vector of auxiliary variables associated with the i-th population element. It is assumed that the sample values of $x_i (i \in s)$, together with the population mean vector $\bar{X} = N^{-1}\sum_{i \in U} x_i$ are known. The vector of sample means is denoted $\bar{x} = n^{-1}\sum_{i \in s} x_i$.

Let y_i denote the value of a survey variable y for the i-th population element and suppose the values of y_i are only observed for $i \in s$. The aim is to estimate the population mean $\bar{Y} = N^{-1}\sum_{i \in U} y_i$.

The regression estimator of $\bar{Y}$ is given by $\bar{y}_r$ in equation (1), where $\bar{y} = n^{-1}\sum_{i \in s} y_i$, $b = \hat{S}_x^{-1}\hat{S}_{xy}$, $\hat{S}_x = n^{-1}\sum_{i \in s}(x_i - \bar{x})(x_i - \bar{x})'$, and $\hat{S}_{xy} = n^{-1}\sum_{i \in s}(x_i - \bar{x})(y_i - \bar{y})$.

This estimator may be motivated by the underlying linear model

$$y_i = \beta_0 + x_i' \beta + \epsilon_i \tag{2}$$

where the ϵ_i are independent disturbances with zero means and common variance σ^2, since we may write $\bar{y}_r = \hat{\beta}_0 + \bar{X}'\hat{\beta}$, where $\hat{\beta}_0 = \bar{y} - \bar{x}' b$ and $\hat{\beta} = b$ are the least squares estimators of β_0 and β, respectively. Under this model the variance of $\bar{y}_r - \bar{Y}$ conditional on the x_i may be written

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}[1 - n/N + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})]. \tag{3}$$

The final term may be interpreted as the effect of estimating β by b. As the number q of x variables increases the residual variance σ^2 may be expected to decrease, but the term $(\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})$ may increase as $\hat{S}_x^{-1}$ becomes more unstable. An alternative way to interpret this term is to write $\bar{y}_r$ as a weighted estimator $\bar{y}_r = n^{-1}\sum_{i \in s} g_i y_i$, where $g_i = 1 + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})$. Then we may write (3) alternatively as

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}(1 - n/N + c_g^2) \tag{4}$$

where c_g is the sample coefficient of variation of the g_i.

To study the expected dependence of c_g^2 on q we now extend the model by supposing that the x_i are independently and identically normally distributed. Noting the independence of $(\bar{x} - \bar{X})$ and $\hat{S}_x$ and also that $E_M(\bar{y}_r - \bar{Y} \mid x_i) = 0$, we obtain the unconditional variance

$$\begin{aligned} &\mathrm{Var}_M(\bar{y}_r - \bar{Y}) \\ &\quad = \sigma^2 n^{-1}\{1 - n/N + \mathrm{tr}[E_M[(\bar{X} - \bar{x})(\bar{X} - \bar{x})']E_M(\hat{S}_x^{-1})]\} \\ &\quad = \sigma^2 n^{-1}(1 - n/N)[1 + q/(n - q - 2)] \end{aligned} \tag{5}$$

using the fact that $n^{-1}\hat{S}_x^{-1}$ has an inverse Wishart distribution (Mardia, Kent and Bibby 1979, p. 69 and 85). This result holds for large n even without normality, in the sense that $[1 - n/N + c_g^2]/(1 - n/N)[1 + q/(n - q - 2)]$ still converges to 1 as n increases for fixed q (under weak conditions).

Expression (5) makes the dependence on q explicit. As q increases we may expect σ^2 to decrease but $E_M(c_g^2)$ to increase. The reduction of σ^2 may be expected to be small after a few important x variables are included and thus the variance may be expected to start increasing at some point where the number of x variables is a nonnegligible fraction of the sample size.

Results (4) and (5) are based on strong modelling assumptions and hence provided us only with motivation. In the general case $\bar{x} - \bar{X} = O_p(n^{-1/2})$ (under the randomization distribution with standard regularity conditions) so that the

last term of (3) is of $O_p(n^{-2})$. A more general second order asymptotic approximation for the design mean squared error of $\bar{y}_r$ when model (2) need not hold may be obtained by generalising Theorem 4.1 of Deng and Wu (1987). Details are given in Silva (1996).

Our aim is to develop a variable selection procedure that minimizes the estimated mean squared error of $\bar{y}_r$, and estimators of this mean squared error are considered next.

3. ESTIMATION OF THE MEAN SQUARED ERROR OF THE MULTIPLE REGRESSION ESTIMATOR

A simple estimator of the mean squared error of $\bar{y}_r$ is obtained by generalizing expression (7.29) of Cochran (1977, p. 195) to the case of several auxiliary variables:

$$v_s = \frac{1-f}{n}\hat{S}_e \tag{6}$$

where $\hat{S}_e = (n-q-1)^{-1}\sum_{i\in s}\hat{e}_i^2$ and $\hat{e}_i = (y_i - \bar{y}) - (x_i - \bar{x})'b$.

This estimator makes no allowance for the $O(n^{-2})$ component of the mean squared error, however. Thus, as a second mean squared error estimator, we generalize the estimator v_d studied in Deng and Wu (1987) to the case of general q. This is a special case of the model-based, bias-robust variance estimator G_2 originally proposed by Royall and Cumberland (1978), for the case where the residual variances in the model (2) are constant. This estimator is given by

$$v_d = \frac{1-f}{n(n-1)}\sum_{i\in s}\alpha_i\hat{e}_i^2 \tag{7}$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f)/\{(1-f)[1-(x_i-\bar{x})'\hat{S}_x^{-1}(x_i-\bar{x})/(n-1)]\}.$$

We originally conjectured that v_d would be second order unbiased, as Deng and Wu (1987, eq. 4.4) show that it is for the case of $q = 1$. However this turns out not to be the case for general $q > 1$, although it may be expected that the bias of v_d is smaller than that of v_s, as indicated by the second order bias expressions for v_s and v_d obtained by Silva (1996).

A difficulty with v_d as a variance estimator is that it does not generalize easily to complex survey designs. Thus we consider as a third variance estimator a modified version of an estimator proposed by Särndal, Swensson and Wretman (1989), defined as:

$$v_g = \frac{1-f}{n(n-q-1)}\sum_{i\in s}g_i^2\hat{e}_i^2. \tag{8}$$

This estimator may be expected to behave similarly to v_d since $\alpha_i = g_i^2 + O_p(n^{-1/2})$. In the terminology of Särndal *et al.* (1992, p. 232), the g_i are the appropriate *g-weights* under simple random sampling if (2) is adopted as the underlying model. Expression (8) differs from the corresponding estimator proposed by Särndal *et al.* (1989, example 4.4) in that we use the denominator $(n - q - 1)$ instead of the original $(n - 1)$.

4. VARIABLE SELECTION PROCEDURES

We consider two basic variable selection procedures. First, an *all subsets* approach that involves computing one of the mean squared error estimators v_s, v_d or v_g of section 3 for all 2^q possible subsets of the q auxiliary variables (always including the intercept) and choosing that subset corresponding to the smallest mean squared error estimate. This procedure can clearly involve considerable computation if q is large. Thus as a second procedure, we consider a *forward selection* approach which starts with the sample mean as an estimator, then adds that variable which minimizes the mean squared error estimate. The procedure is repeated until the mean squared error estimate starts to increase, at which point the subset of variables which gave the minimum mean squared error estimate is selected.

These procedures may be contrasted with an approach inspired by the work of Bankier and his associates – see Bankier (1990) and Bankier, Rathwell and Majkowski (1992). We call this a *condition number reduction approach.* To describe the approach, first note that the regression estimator in (1) can alternatively be expressed as

$$\bar{y}_r = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \tag{9}$$

where X_s^* is the $n\times(q+1)$ matrix with $x_i^{*\prime} = (1, x_{i1}, \ldots, x_{iq})' = (1 \vdots x_i')$ as its i-th row, $\bar{x}^* = (1 \vdots \bar{x}')'$ and $\bar{X}^* = (1 \vdots \bar{X}')'$ are the sample and population mean vectors of x_i^* respectively, and y_s is the $n\times 1$ vector with the sample observations of the response.

The regression estimator thus depends on the inversion of the cross-products matrix $X_s^{*\prime}X_s^*$, a matrix which can sometimes become ill-conditioned and thereby inflate the variance of the regression estimator.

Bankier (1990) proposed a two-step procedure for computing regression estimators of means (or totals) in which columns of the auxiliary data matrix X_s^* were eliminated in order to reduce the condition number of the cross-products matrix $X_s^{*\prime}X_s^*$, as well as to avoid undesirable situations (negative or outlying weights, rare characteristics, or exact linear dependence between columns). Bankier *et al.* (1992) describe in detail the procedure as applied to the 1991 Canadian Population Census. It is worth noting that the approach developed by Bankier and associates, although incorporating variable selection, is not targeted at achieving efficiency for a particular survey variable. Its main focus is on calibration, while at the same time providing a single set of weights that are used for all survey variables.

The condition number reduction approach that we consider can be described by the algorithm below, which adopts a backward elimination procedure to discard auxiliary variables generating large condition numbers for the cross-products matrix $CP = X_s^{*\prime} X_s^*$, instead of the forward inclusion of variables described by Bankier *et al.* (1992).

1) Compute the cross-products matrix $CP = X_s^{*\prime} X_s^*$ considering all the columns initially available (saturated subset).

2) Compute the Hermite canonical form of CP, say H (see Rao 1973, p.18), and check for singularity by looking at the diagonal elements of H. Any zero diagonal elements in H indicate that the corresponding columns of $X_s^{*\prime} X_s^*$ (and X_s^*) are linearly dependent on other columns (see Rao 1973, p. 27). Each of these columns is eliminated by deleting the corresponding rows and columns from $X_s^{*\prime} X_s^*$.

3) After removing any linearly dependent columns, the condition number $c = \lambda_{max}/\lambda_{min}$ of the reduced CP matrix is computed, where λ_{max} and λ_{min} are the largest and smallest of the eigenvalues of CP, respectively. If $c < L$, a specified value, stop and use all the auxiliary variables remaining.

4) Otherwise perform backward elimination as follows. For every k, drop the k-th row and column from CP, and recompute the eigenvalues and the condition number of the reduced matrix. Compute the condition number reductions $r_k = c - c_k$, where c_k is the condition number after dropping the k-th row and column from CP. Determine $r_{max} = \max_k (r_k)$ and $k_{max} = \{k: r_{max} = r_k\}$ and eliminate the column k_{max} by deleting the k_{max} row and column from CP. Make $c = c_{k_{max}}$ and iterate while $c \geq L$ and $q \geq 2$, starting each new iteration with the reduced CP matrix resulting from the previous one.

One further approach that we consider is the ridge regression estimator of Bardsley and Chambers (1984). It does not rely on selecting subsets from the auxiliary variables available, but rather on relaxing the calibration properties of the regression estimator in favour of more stable estimates. The ridge regression estimator is given by

$$\bar{y}_{BC} = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(\lambda C^{-1} + X_s^{*\prime} X_s^*)^{-1} X_s^{*\prime} y_s]/N \tag{10}$$

where λ is a scalar ridging parameter and C is a diagonal matrix of "cost" coefficients associated with the calibration errors tolerated when estimating totals of the auxiliary variables using $\bar{y}_{BC}$.

Bardsley and Chambers (1984) suggested that the specification of the matrix C could be used to control the influence of each auxiliary variable on the resulting estimator of the response mean, thus imitating the subset selection process. As for the ridging parameter λ, they suggested taking the smallest value such that all the implicit case weights are not smaller than $1/N$ (or 1 for estimating totals).

5. PROPERTIES OF REGRESSION ESTIMATORS AFTER VARIABLE SELECTION

For our basic variable selection procedures, a set of estimation strategies $\boldsymbol{S} = \{(\bar{y}_r^\gamma, v^\gamma); \gamma \in \Gamma\}$ is considered, where $\bar{y}_r^\gamma$ and v^γ are the regression estimator and an estimator of its variance respectively for a subset γ of the q auxiliary variables available, and Γ is the set of all subsets. The variable selection procedure selects a subset γ^* from Γ according to a rule which is determined by the data and by $\boldsymbol{S}$, and the resulting point estimator is $\bar{y}_r^{\gamma^*}$.

For each fixed subset γ, it follows under standard regularity conditions (Isaki and Fuller 1982) that $\bar{y}_r^\gamma$ is consistent for the population mean $\bar{Y}$, that is $\bar{y}_r^\gamma - \bar{Y} = o_p(1)$. Now, for given $\delta > 0$, $|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta$ implies $|\bar{y}_r^\gamma - \bar{Y}| > \delta$ for some γ, and so we have

$$\Pr(|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta) \leq \sum_{\gamma \in \Gamma} \Pr(|\bar{y}_r^\gamma - \bar{Y}| > \delta) \tag{11}$$

and because Γ is finite, the right hand side of (11) converges to zero, and it follows that $\bar{y}_r^{\gamma^*}$ is also consistent.

The distribution of $\bar{y}_r^{\gamma^*}$ will, however, depend on the selection rule in a complex way. See Grimes and Sukhatme (1980) for an investigation of the efficiency of $\bar{y}_r^{\gamma^*}$ in the simplest case when there are just two possible estimators: a regression estimator with one x variable and a difference estimator (a special case of which is the mean) and the variables are jointly normally distributed.

In contrast to the consistency of $\bar{y}_r^{\gamma^*}$, there is no reason why v^{γ^*} should be consistent for $\mathrm{Var}(\bar{y}_r^{\gamma^*})$, even if v^γ is consistent for $\mathrm{Var}(\bar{y}_r^\gamma)$ for each fixed γ. In particular we may expect v^{γ^*} to underestimate $\mathrm{Var}(\bar{y}_r^{\gamma^*})$ if the selection rule is such that v^{γ^*} is the minimum of the v^γ. This effect is similar to the well known overestimation of R^2 after subset selection in standard multiple linear regression (Miller 1990, p. 7-10).

6. A SIMULATION STUDY

In this section we present a small simulation study carried out to evaluate the performance of the alternative variable selection procedures considered. We took as our simulation population a data set comprising 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil.

This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. The test consisted of two rounds of data collection. In the first round, each enumerator would visit all the occupied households in a given enumeration area (an area with between 200 and 300 households on average) and would fill in a short questionnaire. This form contained a few questions about characteristics of the household and about each member of the household (sex, age, relationship to head of household

and literacy). For heads of household only, a question on education and another about monthly total income were also included. The reported monthly total income for heads of household provides only a proxy to the actual income, due to the limitations of the interviewing process in this first round of data collection.

Then a second round of data collection was undertaken in each enumeration area. The same enumerators would visit a sample of 1 in 10 of the households (selected systematically from the list of occupied households compiled in the first round of data collection) to obtain information using a long (more detailed) questionnaire, which contained all the questions asked in the short form plus many other questions.

The size of the surveyed population was approximately 44,000 households with 188,000 individuals. The sample size was roughly 10% of the population size. For reasons of computational cost, we used in our simulation study a sub-population comprising all the sample records for 426 heads of household living in 20 of the 170 enumeration areas. We chose these records as our simulation population because they contain all the detailed information provided in the sample questionnaire, as well as the proxy information available from the first round interviews using the short form.

We considered total monthly income, as obtained from the long form, as the main response variable (y) together with 11 potential auxiliary variables, namely:

x_1 = indicator of sex of head of household equal male;
x_2 = indicator of age of head of household less than or equal to 35;
x_3 = indicator of age of head of household greater than 35 and less than or equal to 55;
x_4 = total number of rooms in household;
x_5 = total number of bathrooms in household;
x_6 = indicator of ownership of household;
x_7 = indicator that household type is house;
x_8 = indicator of ownership of at least one car in household;
x_9 = indicator of ownership of colour TV in household;
x_{10} = years of study of head of household;
x_{11} = proxy of total monthly income of head of household.

From these 11 variables, we constructed two alternative sets of auxiliary variables for our simulations. The first set was defined by taking five auxiliary variables, namely $x_1,\ldots,x_4$ and x_{11}, that have reasonable explanatory power in predicting y, especially due to the presence of the proxy income x_{11}. The second set we considered contained ten auxiliary variables, namely $x_1,\ldots,x_{10}$, which due to the exclusion of x_{11}, has smaller predictive power than the previous one. For reference, the population correlation matrix for the survey variable y and the 11 auxiliary variables in the population is given in Table 3.

We then selected 1,000 samples of size 100 from this simulation population by simple random sampling without replacement.

Before proceeding to examine the detailed simulation results, we first consider the potential for gains from variable selection following the motivating model-based discussion of section 2. Recall from equation (4) that under model (2) the conditional variance of $\bar{y}_r$ is inflated by a term c_g^2 because of estimation of β. We evaluated the distribution of c_g^2 over the 1,000 samples for both the cases of five and ten auxiliary variables. For the case of five auxiliary variables, the median value of c_g^2 was 0.036, with upper quartile of 0.056 and maximum 0.255. This accords roughly with equation (5) which implies that under the model the expected value of c_g^2 is $(1 - n/N)q/(n - q - 2) = 0.041$. Note that the wide variation of c_g^2 across samples suggests that it may be sensible to adopt a procedure which selects a different set of variables for each sample. The variation of c_g^2 is even greater for the case of ten auxiliary variables, when the median was 0.078, the upper quartile was 0.107 and the maximum was 0.329, which also accords roughly with the expected value under the model of 0.087, according to equation (5). This interpretation clearly depends on the validity of the model (2), which is doubtful for these data, but it does suggest that there are potential efficiency gains to be made from variable selection.

Another way to assess the potential for efficiency gains from variable selection is to compute approximations to the variance of the regression estimator considering various subsets of the auxiliary variables available, using all the population records. Figure 1 displays a plot of the approximation given by a finite population version of equation (5) computed for increasing subsets of the ten auxiliary variables, where the variable added at each step is the one yielding the biggest decrease in the approximation. The values of the standard first order design-based approximation $(1 - f)S_e/n$ are also plotted for reference, although as has already been noted, this approximation is monotone non-increasing when new auxiliary variables are added. Simulation estimates of the mean squared error for the regression estimator corresponding to each subset are also plotted. The plot shows clearly that if a standard regression estimator with a fixed set of auxiliary variables is to be used, the subset with five predictors would be the best choice when

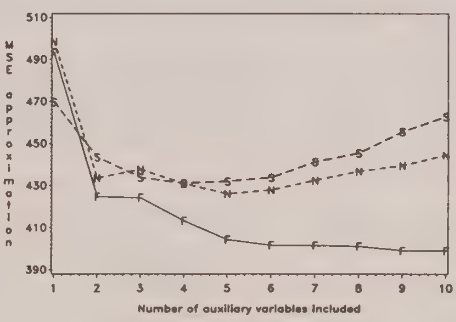


Figure 1. Finite population approximations and simulation estimations for the MSSE of the regression estimator computed for increasing subsets of the ten auxiliary variables.

the normal approximation for the variance based on expression (5) was considered, whereas the saturated subset would be chosen in case the standard design-based approximation for the variance was considered. The plot also reveals that the simulation estimates of the mean squared error agree more closely with the normal model approximation than with the standard first order approximation, especially for larger subsets of auxiliary variables. Similar results are achieved when corresponding variance approximations are computed given the set of five auxiliary variables.

Hence both the simulation distributions of c_g^2 and the finite population approximations to the variance of the regression estimator indicate that there are potential efficiency gains to be made from variable selection for this population. To investigate this for our data we now proceed to describe the details of the simulation study.

For each sample replicate (say s) and for each of the two alternative sets of auxiliary variables considered, estimates of the population mean of total monthly income were computed, as well as corresponding variance estimates, using a number of estimation strategies. Each estimation strategy is defined as a combination of a subset selection procedure, an estimator for the mean and a corresponding variance estimator. The list of all strategies considered follows.

SM) Sample mean estimator, with no auxiliary variables $(\bar{y}, v_s)$. This strategy provides the standard against which all the others will be compared.
Fs) Forward selection of auxiliary variables with $(\bar{y}_r, v_s)$.
Fd) Forward selection of auxiliary variables with $(\bar{y}_r, v_d)$.
Fg) Forward selection of auxiliary variables with $(\bar{y}_r, v_g)$.
Bs) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_s)$.
Bd) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_d)$.
Bg) Best subset selection from all subsets of auxiliary variables with $(\bar{y}_r, v_g)$.
FI) Fixed subset of auxiliary variables with $(\bar{y}_r, v_s)$.
SS) Saturated subset of auxiliary variables with $(\bar{y}_r, v_s)$.
FR) Forward subset selection using SAS PROC REG, with $(\bar{y}_r, v_s)$.
CN) Condition number reduction subset selection procedure with $(\bar{y}_r, v_s)$.
RI) Ridge regression estimator with saturated subset of auxiliary variables and a variance estimator that we denote v_{DC}, proposed by Dunstan and Chambers (1986), $(\bar{y}_{BC}, v_{DC})$.

Strategies Fs to Bg are variations of the two procedures we proposed for subset selection arising from the use of the three mean squared error estimators considered in section 3. Strategies FI and SS use the same set of auxiliary variables irrespective of the sample selected. In SS the saturated subset including all auxiliary variables available is always used. In FI a subset was chosen from each of the two sets with five (x_1, x_4, x_{11} chosen) or ten ($x_1, x_2, x_5, x_8, x_{10}$ chosen) auxiliary variables considered, by applying a standard forward subset selection regression procedure to the population dataset. The selected subsets were then used for every sample, thus the name "fixed subset" strategy for FI. This strategy would not be feasible in practice because the population information would not be available for the response, but it was considered as a theoretical "best possible scenario" under the traditional approach.

For the strategy FR, SAS PROC REG was used "naively" to perform a standard forward subset selection for each sample. The *p-value* used to decide whether a new variable should be included was the default of the procedure, namely 0.50. For more details, see SAS (1990, p. 1397).

For the condition number reduction subset selection strategy CN, the value used for the parameter L that controls the method was 1,000. For the ridge regression estimator strategy RI, the cost coefficients associated with calibration errors for different variables were all set equal to 1. After having chosen the value of λ that guarantees all the weights are not less than $1/N$, the weights were rescaled such that they sum to exactly 1, in order to ensure exact calibration when estimating the population size.

For any estimation strategy, the estimates of the population mean and its mean squared error for the sample s are denoted by $\bar{y}(s)$ and $v[\bar{y}(s)]$ respectively. The simulation results for each estimation strategy were summarised by computing estimates of the bias, mean squared error (MSE), and average of mean squared error estimates (AVMSE) from the set of 1,000 sample replicates, given respectively by

$$\text{BIAS} = \sum_s [\bar{y}(s) - \bar{Y}]/1{,}000 \tag{12}$$

$$\text{MSE} = \sum_s [\bar{y}(s) - \bar{Y}]^2/1{,}000 \tag{13}$$

$$\text{AVMSE} = \sum_s v[\bar{y}(s)]/1{,}000. \tag{14}$$

A measure of efficiency was also calculated for each strategy by dividing the corresponding simulation mean squared error by the simulation mean squared error for the sample mean (strategy SM) and multiplying the result by 100. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each estimation strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 2.

Table 1 displays the simulation results for estimation of the population mean of the response variable given the set of five auxiliary variables ($x_1 - x_4, x_{11}$) with larger predictive power. In this case, the use of the regression estimator greatly improves precision for every estimation strategy employed, except for subset selection using condition number reduction (CN). The bias was negligible (less than 1% in terms of the absolute relative bias) for all estimation strategies (the population mean of y is 194.34) except perhaps RI, which displayed a slight bias.

Table 1
Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Five Auxiliary Variables ($x_1 - x_4, x_{11}$) Available

Estimation strategy	BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM) Sample mean $(\bar{y}, v_s)$	0.25	620.09	619.05	100.00	91.8
Fs) Forward $(\bar{y}_r, v_s)$	0.40	233.78	239.62	37.70	82.7
Fd) Forward $(\bar{y}_r, v_d)$	-1.25	188.08	196.88	30.33	82.0
Fg) Forward $(\bar{y}_r, v_g)$	-1.28	188.38	192.73	30.38	81.1
Bs) Best $(\bar{y}_r, v_s)$	0.44	236.90	239.49	38.20	82.7
Bd) Best $(\bar{y}_r, v_d)$	-1.22	190.52	196.84	30.72	82.0
Bg) Best $(\bar{y}_r, v_g)$	-1.24	190.83	192.71	30.77	81.1
FI) Fixed $(\bar{y}_r, v_s)$	0.29	227.90	241.24	36.75	83.3
SS) Saturated $(\bar{y}_r, v_s)$	0.30	233.58	242.32	37.67	82.5
FR) PROC REG $(\bar{y}_r, v_s)$	0.38	235.86	240.26	38.04	82.5
CN) Cond. num. red. $(\bar{y}_r, v_s)$	0.34	507.33	483.63	81.82	89.8
RI) Ridge $(\bar{y}_{BC}, v_{DC})$	2.12	304.95	257.07	49.18	82.5

[1] Nominal 95% coverage.

There was no difference between the results for strategies based on forward selection (Fs-Fg) and corresponding strategies based on selection from all possible subsets (Bs-Bg). Hence the faster and cheaper forward selection procedures are preferable.

Amongst the strategies using forward subset selection, Fd and Fg (with v_d and v_g as the mean squared error estimators respectively) yielded greater efficiency, and performed very similarly. Note also that Fd and Fg performed better than FI and SS, the strategies that adopted the regression estimator with a fixed subset of the five auxiliary variables for every sample. This is true both for the saturated subset (SS) and when the fixed subset was chosen using information from the whole population (FI). This shows that one can do better than the traditional approach of using the regression estimator with a fixed set of auxiliary variables, by using an adaptive procedure that chooses the "best" regression estimator (subset) for each given sample, at least when the target response variable is the one considered for subset selection. This property was suggested by the wide variation in the values of c_g^2 between samples, where we may expect to benefit from a strategy which selects fewer x variables for samples with the largest values of c_g^2.

Comparison with the adaptive strategy FR, which used the standard subset selection available in PROC REG of SAS, shows that a criterion using an appropriate estimator of the mean squared error of the regression estimator makes some difference. FR yielded similar efficiency to that of traditional fixed subset strategies (FI-SS).

A more striking result is the low efficiency achieved by the subset selection procedure based on condition number reduction (CN) compared to all the other strategies based on the regression estimator. This was not unexpected, because that procedure did not take the response variable into account. This favours the argument that when the mean of some specified response variable is the main target for inference, this should be taken into account when selecting the auxiliary variables to use in connection with the regression estimator.

When the set of five auxiliary variables was considered, we also observed that, for every sample, the first variable eliminated to reduce the condition number was proxy income (x_{11}). This happened because eigenvalues (and hence condition numbers) of the CP matrix are dependent on the units of measurement of the auxiliary variables. Because all other auxiliary variables are counts of some kind, proxy income is the variable with the largest variance by far. Its exclusion for every sample provides some explanation for the poor performance of this approach, because it is the best single predictor for the response.

This difficulty was not apparent in Bankier's work, because in the target application of his procedure, the sample data from the 1991 Canadian Population Census, all the auxiliary variables considered were counts of persons, families or households, thus measured in similar units.

Unlike the eigenvalues of the CP matrix, the regression estimator is invariant to location and scale transformation of the auxiliary variables. To remove the arbitrary dependence of the condition number approach on the units of the auxiliary variables, it is therefore natural to standardise these variables first and to compute the condition number of the sample correlation matrix $\hat{R}_x$ rather than $X_s^{*\prime} X_s^*$. However this was tried and even modest values of L (100) failed to cause elimination of any auxiliary variables, which resulted in the saturated set being used every time, so that CN reduced to SS.

The strategy based on the ridge regression estimator (RI) performed worse than the saturated subset strategy (SS) in terms of efficiency. It also displayed some bias for estimating the mean squared error. This loss of efficiency is due to the

requirement that all the weights should be greater than or equal to $1/N$, which was imposed only under this strategy. On the other hand, it performed much better than the condition number reduction strategy CN in terms of efficiency.

In terms of the empirical coverage rates, only the condition number reduction strategy CN performed close to SM (sample mean), both leading to modest undercoverage. All the other strategies based on regression estimation yielded similar coverage rates, well below the target of 95%.

Results for the simulation carried out with the set of ten auxiliary variables ($x_1 - x_{10}$) are displayed in Table 2 below. As expected, these results show that the strategies that use the regression estimator still provide some gain in efficiency over the sample mean. However these gains are not as large as those reported in Table 1, when there are five auxiliary variables with higher explanatory power. As before, adaptive strategies based on forward subset selection performed similarly to their counterparts based on best subset selection from all possible subsets. Adaptive strategies using v_d or v_g as the estimator of the mean squared error were again slightly more efficient than the corresponding strategies based on v_s, although in this case at the expense of larger undercoverage of the corresponding nominal 95% confidence intervals.

The more efficient adaptive estimation strategies (Fd, Fg, Bd and Bg) display nonnegligible bias for both the population mean and for the mean squared error. In contrast, strategies FI and SS present no significant bias for the mean, although there is some bias in the mean squared error estimation under strategy SS. Note particularly the large negative bias of the estimators of the mean squared error, as indicated by the differences between the columns labelled MSE and AVMSE in Table 2. This appears to be worse for strategies Fd, Fg, Bd and Bg, followed by Fs and Bs, and not so bad for SS, FR and CN.

Comparing Fd and Fg with CN, there is a moderate gain in efficiency over the condition number reduction procedure, at the expense of some increased bias in both the mean and mean squared error estimators. Thus, even when the predictive power of the available auxiliary variables is not large, it is still possible to gain efficiency over strategy CN.

A bad choice of fixed subset (as for example, the saturated subset used in strategy SS) could yield poor results in terms of efficiency and also some bias in the mean squared error estimation. However, if for example v_d was used as the estimator for the mean squared error under strategy SS instead of v_s, there would be no apparent bias (the AVMSE observed in that case was 459.67, hence much closer to the estimated simulation mean squared error of 462.71).

The ridge regression estimator was again slightly inferior to the saturated subset strategy (SS), but now without any apparent bias in estimating the mean or the mean squared error. It outperformed the condition number reduction strategy CN once again in terms of efficiency, albeit by a smaller margin. It also performed well in terms of empirical coverage.

Strategy FR performed similarly to the fixed subset strategies FI and SS again, and so was outperformed by strategies using a specialized criterion based on an estimator of the mean squared error of the regression estimator such as v_d or v_g.

These results suggest that, when estimating the population mean of a single response, the proposed adaptive procedures combining the regression estimator with some form of subset selection based on an appropriate mean squared error estimator can offer some useful improvements in efficiency against its competitors. However such strategies may introduce some bias when the predictive power of the auxiliary variables available is not large, and the corresponding MSE estimators may be substantially biased, leading to poor coverage.

Table 2
Bias, Mean Squared Error, Average of Mean Squared Error Estimates, Efficiency and Empirical Coverage of Alternative Estimation Strategies for the Mean of Response Variable y with Ten Auxiliary Variables (x_1–x_{10}) Available

Estimation strategy	BIAS	MSE	AVMSE	Efficiency over SM (%)	Empirical[1] Coverage (%)
SM) Sample mean $(\bar{y}, v_s)$	0.25	620.09	619.05	100.00	91.8
Fs) Forward $(\bar{y}_r, v_s)$	0.06	468.46	397.99	75.55	86.7
Fd) Forward $(\bar{y}_r, v_d)$	-8.12	434.27	338.90	70.03	81.7
Fg) Forward $(\bar{y}_r, v_g)$	-7.90	433.71	328.46	69.94	81.6
Bs) Best $(\bar{y}_r, v_s)$	-0.00	466.16	397.59	75.18	86.6
Bd) Best $(\bar{y}_r, v_d)$	-7.90	434.54	336.88	70.08	81.5
Bg) Best $(\bar{y}_r, v_g)$	-7.60	433.26	326.05	69.87	81.6
FI) Fixed $(\bar{y}_r, v_s)$	0.45	490.49	461.86	79.10	89.0
SS) Saturated $(\bar{y}_r, v_s)$	-0.20	462.71	413.17	74.62	86.9
FR) PROC REG $(\bar{y}_r, v_s)$	-0.07	466.13	399.34	75.17	86.4
CN) Cond. num. red. $(\bar{y}_r, v_s)$	3.49	562.91	450.36	90.78	87.3
RI) Ridge $(\bar{y}_{BC}, v_{DC})$	1.05	480.18	472.82	77.44	89.4

[1] Nominal 95% coverage.

Table 3
Correlation Matrix for Variables Used in the Simulation Study with the 1988 Census Population

Variable	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	0.23										
x_2	-0.04	0.20									
x_3	0.17	0.07	-0.40								
x_4	0.47	0.13	-0.15	0.12							
x_5	0.48	0.09	-0.11	0.15	0.83						
x_6	0.05	-0.09	-0.32	-0.03	0.22	0.20					
x_7	-0.17	0.01	-0.12	-0.01	-0.17	-0.31	0.16				
x_8	0.38	0.29	0.07	0.17	0.44	0.41	0.13	-0.20			
x_9	0.20	0.08	-0.06	0.04	0.30	0.25	0.16	-0.13	0.37		
x_{10}	0.43	0.23	0.33	0.17	0.39	0.39	-0.10	-0.30	0.49	0.26	
x_{11}	0.78	0.23	-0.00	0.22	0.54	0.54	0.01	-0.19	0.41	0.21	0.49

7. CONCLUSIONS AND FUTURE DIRECTIONS

Our results suggest that, when using regression estimation, there is potential for some gain in efficiency by adopting a variable selection procedure based on one of the mean squared error estimators v_d or v_g. Under SRS, and considering the limited simulation evidence, there seems little to choose between these two mean squared error estimators.

Forward subset selection procedures were as effective as those based on searches carried out considering all possible subsets, which involve much more computation. Our results also indicate that it is possible to improve over subset selection procedures based on condition number reduction whenever a specific response variable is of interest.

One problem with a variable selection approach is that the associated variance estimation is likely to become biased for the estimation of the overall mean squared error of the regression estimator following variable selection, thus leading to poor coverage of standard confidence interval procedures. Further research is necessary to investigate possible alternative variance estimation procedures.

This paper has focused on the use of regression estimation to reduce sampling variance in the classical sampling context. In practice, regression estimation is widely used to correct for biases arising from non-sampling errors. In such applications the question of how many auxiliary variables to use is also an important one. Some variables might be included for reasons unrelated to sampling error, for example because they are known to be important determinants of nonresponse. Nevertheless, as the number of auxiliary variables increases the sampling variance may also eventually increase and we suggest that a decision rule to limit the number of auxiliary variables employed might still usefully be based on sampling variance considerations. In the presence of nonsampling bias, the difference between $\bar{x}$ and $\bar{X}$ will generally be of $O_p(1)$ not $O_p(n^{-1/2})$ and so the results of this paper are not directly applicable. Further research is therefore needed to consider the extension of our approach to this case.

Further research is also necessary to extend our approach to complex sampling designs. One possible approach for the general regression estimators, considered *e.g.* by Särndal *et al.* (1992, sec. 6.4), would be to replace the weights g_i by the "generalized" weights, described by Särndal *et al.* (1992, eq. 6.5.9), and to base variable selection on the minimization of the generalized version of v_g given by Särndal *et al.* (1992, eq. 6.6.4).

ACKNOWLEDGEMENTS

Pedro L.D. Nascimento Silva is grateful to CVCP-UK, CNPq-Brasil and IBGE-Brasil for financial support. The authors are grateful to Ray Chambers, Danny Pfeffermann, Jon Rao, Michael Bankier and two anonymous referees for comments. Michael Bankier was also very helpful for providing documentation and software about his GLSEP procedure.

REFERENCES

BANKIER, M.D. (1990). Two Step Generalized Least Squares Estimation. Ottawa: Statistics Canada, Social Survey Methods Division, internal report.

BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Methodology Branch Working Paper, SSMD, 92-007E, Statistics Canada.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in mltipurpose surveys. *Applied Statistics*, 35, 276-280.

GRIMES, J.E., and SUKHATME, B.V. (1980). A regression-type estimator based on preliminary test of significance. *Journal of the American Statistical Association*, 75, 957-962.

HANSEN, M.H., and TEPPING, B.J. (1969). Progress and problems in survey methods and theory illustrated by the work of the United States Bureau of the Census. *New Developments in Survey Sampling*, (N.L. Johnson and H. Smith Jr., Eds.). New York: John Wiley & Sons.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MILLER, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). New York: John Wiley & Sons.

ROYALL, R.M., and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide* (Version 6, Vol. 2, 4th ed.). Cary, NC: SAS Institute Inc.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SILVA, P.L.D.N. (1996). Some Asymptotic Results on the Mean Squared Error of the Regression Estimator Under Simple Random Sampling Without Replacement. Southampton: University of Southampton, Centre for Survey Data Analysis Technical Report 96-2.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 33-40
Statistics Canada

Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey

JOHN L. ELTINGE and IBRAHIM S. YANSANEH[1]

ABSTRACT

This paper discusses the use of some simple diagnostics to guide the formation of nonresponse adjustment cells. Following Little (1986), we consider construction of adjustment cells by grouping sample units according to their estimated response probabilities or estimated survey items. Four issues receive principal attention: assessment of the sensitivity of adjusted mean estimates to changes in k, the number of cells used; identification of specific cells that require additional refinement; comparison of adjusted and unadjusted mean estimates; and comparison of estimation results from estimated-probability and estimated-item based cells. The proposed methods are motivated and illustrated with an application involving estimation of mean consumer unit income from the U.S. Consumer Expenditure Survey.

KEY WORDS: Incomplete data; Missing data; Quasi-randomization; Response propensity; Sensitivity analysis; Weighting adjustment.

1. INTRODUCTION

1.1 Problem Statement

Survey analysts often use adjustment cell methods to account for nonresponse. The main idea is to define groups, or "cells", of sample units which are believed to have approximately equal response probabilities, or approximately equal values of a specific survey item, *e.g.*, income. Weighting adjustment or simple hot-deck imputation then is carried out separately within each adjustment cell. The resulting adjusted estimator of a population mean or total will have a nonresponse bias approximately equal to zero, provided the within-cell covariances between survey items and response probabilities are approximately equal to zero.

Some previous nonresponse-adjustment work formed adjustment cells through combinations of simple demographic or geographical classificatory variables. However, Little (1986) and others considered formation of cells by direct grouping of sample units according to their estimated response probabilities or estimated item values. The present paper discusses some simple diagnostics that are useful in implementing these cell-formation ideas. Principal attention is directed to the sensitivity of results to the number of cells used; identification of specific cells that require additional refinement; comparison of adjusted and unadjusted mean estimates; and comparison of estimation results from estimated-probability and estimated-item based cells. These diagnostics are illustrated with income data collected in the U.S. Consumer Expenditure Survey.

1.2 Notation, Nonresponse Bias, and Adjustment Cells

Let U be a fixed population of size N with survey items Y_i, $i \in U$; and consider estimation of the population mean $\bar{Y} = N^{-1}\sum_{i\in U} Y_i$. A sample s of size n is selected from U, and π_i is the probability that unit i is included in the sample.

Nonresponse is assumed to satisfy the following quasi-randomization model (Oh and Scheuren 1983). Let R_i be an indicator variable equal to 1 if the selected sample unit i is a respondent and equal to 0 otherwise. Assume that the R_i are mutually independent Bernoulli (η_i) random variables, where the fixed response probabilities η_i are allowed to differ across units. In addition, define the survey weights $\lambda_i = \pi_i^{-1}$ and the unadjusted survey-weighted mean response

$$\hat{\bar{Y}}_1 \stackrel{\text{def}}{=} \left(\sum_{i\in s} \lambda_i R_i\right)^{-1} \sum_{i\in s} \lambda_i R_i Y_i. \tag{1.1}$$

Because of differences among the η_i, the unadjusted estimator $\hat{\bar{Y}}_1$ has a nonresponse bias approximately equal to $N^{-1}\bar{\eta}^{-1}\sum_{i\in U}\eta_i(Y_i - \bar{Y})$, where $\bar{\eta} = N^{-1}\sum_{i\in U}\eta_i$ and expectations are taken over both the original sample design and the quasi-randomization model. To reduce this bias, one often partitions the population into k "adjustment cells" U_h, partitions the sample s into corresponding groups s_h, and then uses the adjusted estimator

$$\hat{\bar{Y}}_k \stackrel{\text{def}}{=} \sum_{h=1}^{k} w_h \bar{Y}_{hR}, \tag{1.2}$$

where $w_h = (\sum_{i\in s}\lambda_i)^{-1}\sum_{i\in s_h}\lambda_i$ and $\bar{Y}_{hR} = (\sum_{i\in s_h}\lambda_i R_i)^{-1} \sum_{i\in s_h}\lambda_i R_i Y_i$. Note that if $k = 1$, then estimators (1.1) and (1.2) are identical. For some general discussion of adjustment cell methods see, *e.g.*, Cassel, Särndal and Wretman (1983), Oh and Scheuren (1983), and Kalton and Maligalig (1991).

The adjusted estimator $\hat{\bar{Y}}_k$ has remaining nonresponse bias approximately equal to

$$N^{-1}\sum_{h=1}^{k} \bar{\eta}_h^{-1} \sum_{i\in U_h} (\eta_i - \bar{\eta}_h)(Y_i - \bar{Y}_h), \tag{1.3}$$

[1] John L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.; Ibrahim S. Yansaneh, Westat, 1650 Research Blvd., Rockville, MD 20850-3195, U.S.A.

where N_h is the number of units in U_h and $(\bar{\eta}_h, \bar{Y}_h) = N_h^{-1} \sum_{i \in U_h} (\eta_i, Y_i)$. Consequently, one prefers to construct cells such that the population covariance between η_i and Y_i is approximately equal to zero within each cell. In practice, one attempts to accomplish this by constructing cells that are approximately homogeneous in the response probabilities η_i or in the items Y_i, or both. In some cases, "natural" sets of cells are defined *a priori* through combinations of classificatory variables that are available for both respondents and nonrespondents. For example, Ezzati and Khare (1992) used 72 cells defined by age, race, region, urbanization status, and household size to perform nonresponse adjustments for part of the National Health and Nutrition Examination Survey. In many practical cases, however, the list of reasonable candidate variables for cell formation is fairly large, and may produce a substantial number of cells that contain few, if any, respondents. Consequently, several authors have developed methods to screen out the less important classificatory variables and to collapse sparse adjustment cells in a way that preserves a reasonable degree of homogeneity within each of the remaining cells. See, *e.g.*, Tremblay (1986); Lepkowski, Kalton and Kasprzyk (1989); Kalton and Maligalig (1991); Göskel, Judkins and Mosher (1991); and the related discussion of pooling of poststrata in Little (1993). In addition, adjustment cell methods are related to other methods like regression-based adjustments (*e.g.*, Rao 1996, Section 2.4 and references cited therein) and generalized raking (Deville, Särndal and Sautory 1993).

1.3 Adjustment Cells Based on Estimated Response Propensities or Predicted Items

Adjustment cells are expected to be approximately homogeneous, so one may argue that such cells implicitly define a model for either the η_i or Y_i values, or both. More explicit modeling leads to two related cell formation methods. First, let X_i be a vector of auxiliary variables observed for both responding and nonresponding sample units i, and use the sample (R_i, X_i) values to fit a model for $\eta_i = \eta(X_i)$ through linear, logistic, or probit regression. The sample cells s_h are then formed by grouping the sample units according to their estimated response probabilities $\hat{\eta}_i$. As a second alternative, consider regression of responses Y_i on an auxiliary vector X_i to produce estimated items $\hat{Y}_i$ for both responding and nonresponding sample units. The sample cells s_h are then formed by grouping units according to the values $\hat{Y}_i$.

These two methods were suggested by Little (1986), extending the observational-data propensity-score work of Rosenbaum and Rubin (1983, 1984). See also David, Little, Samuhel and Triest (1983). These ideas were developed originally in a model- based context, but extend directly to the current framework. Little (1986) argued that use of cells based on either the $\hat{\eta}_i$ or $\hat{Y}_i$ values could reduce nonresponse bias, and that the $\hat{Y}_i$-based cells could also control variance. Also, in some cases the $\hat{\eta}_i$ and $\hat{Y}_i$-based cells can be more flexible than cells defined *a priori*. In addition, the $\hat{Y}_i$-based adjustment cells are conceptually related to optimum stratification ideas (*e.g.*, Cochran 1977, Sections 5A.7-5A.8).

Little (1986) did not propose a specific rule to determine cell divisions. However, in keeping with related observational-data work by Cochran (1968) and by Rosenbaum and Rubin (1984), one may consider cell divisions defined by the estimated $k^{-1}j$ quantiles of the $\hat{\eta}_i$ or $\hat{Y}_i$ populations, $j = 1, 2, ..., k-1$. This equal-quantile method gives some control over the expected number of respondents in each cell. In addition, review of the preceding two references suggests that, for a given set of predictors X_i, most of the feasible bias reduction may be achieved with a relatively small number of cells, say $k = 5$. A case study by Czajka, Hirabayashi, Little and Rubin (1992) used $k = 6$ $\hat{\eta}_i$-based adjustment cells within each of several strata, using cell-formation rules that were somewhat more complex than the equal-quantile rule considered here. However, the potential adequacy of a small number of cells should not be over-interpreted. For example, if an important regressor is omitted, then the resulting cell-based adjusted estimators may retain a substantial amount of bias, regardless of the specific number of estimated-probability or estimated-item based cells used.

Finally, an important alternative to weighting adjustment is imputation. For example, simple hot-deck imputation replaces a missing value within a given adjustment cell by randomly selecting respondent donors from the same cell. In parallel with (1.1) and (1.2), the resulting mean estimator is $\hat{\bar{Y}}_{\text{imp}} = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s} \lambda_i Y_i^*$, where Y_i^* is either an observed or imputed value, as appropriate. Practical applications often use weighting adjustment for unit nonresponse and imputation for item nonresponse. However, for a given set of cells, both the weighting adjustment point estimator (1.2) and the imputation estimator $\hat{\bar{Y}}_{\text{imp}}$ have the same approximate bias (1.3). For simplicity, the remainder of this paper will focus on weighting adjustment, but one should bear in mind that for a given set of cells, the same bias-reduction issues arise regardless of whether those cells are used for weighting adjustment or simple hot deck imputation.

1.4 Outline of the Present Paper

This paper discusses some implementation details of the estimated-probability and estimated-item methods of cell formation. We devote special attention to diagnostics to identify problems in a specific set of cells, and motivate and illustrate these diagnostics with an extended example involving income nonresponse in the U.S. Consumer Expenditure Survey. Section 2 gives some general background on this income nonresponse problem. Section 3 describes and applies several diagnostics, including comparison of $\hat{\bar{Y}}_k$ estimates and standard errors for several values of k (Section 3.1); partial assessment of within-cell bias (Section 3.2.1); assessment of cell widths relative to the precision of $\hat{\eta}_i$ estimates (Section 3.2.2); and comparison of the adjusted and unadjusted mean estimates $\hat{\bar{Y}}_k$ and $\hat{\bar{Y}}_1$ (Section 3.3). Section 4 shows that similar diagnostics can be applied to adjustment cells based on predicted incomes $\hat{Y}_i$,

and also compares the mean income estimates computed from estimated-probability and estimated-income based cells. Section 5 summarizes the main ideas used in this paper, and notes some areas for future research.

2. INCOME NONRESPONSE IN THE U.S. CONSUMER EXPENDITURE SURVEY

2.1 The Consumer Expenditure Survey, Weighting Methods and Variance Estimation

The U.S. Consumer Expenditure Survey (CE) is a stratified multistage rotation sample survey conducted by the Census Bureau for the Bureau of Labor Statistics. Sample elements are "consumer units", roughly equivalent to households. In the interview component of this survey, each selected sample unit is asked to participate in five interviews. The current CE weighting procedure accounts for initial selection probabilities, a noninterview adjustment, post-stratification based on several demographic variables, and additional refinements; see Zieschang (1990) and United States Bureau of Labor Statistics (1992). The complexity of the CE weighting work has led the BLS to use variance estimators based on pseudo-replication methods with 44 replicates. This pseudo-replication is approximately equivalent to standard balanced repeated replication (Wolter 1985, Ch. 3). All standard errors reported here are based on this pseudo-replication method, with all additional parameter estimation and weighting adjustment steps performed separately within each replicate.

2.2 Income Nonresponse

The noninterview adjustment in the current CE weighting procedure is generally considered to account adequately for unit nonresponse, *e.g.*, noncontact or refusal to participate in a specific interview. Thus, unit nonresponse in the CE will not be considered further here. However, the BLS has had concerns about possible bias in mean income estimates due to item nonresponse that occurs with income questions in the CE; some background is as follows.

Detailed income data are collected in the second and fifth interviews of the CE, and are used to produce estimates of mean consumer unit income (U.S. Bureau of Labor Statistics 1991) and other parameters. CE income data are collected through a complex set of questions, and nonresponse rates for these questions are relatively high. To provide a summary indication of response or nonresponse to the full set of income questions, the BLS classifies each second- or fifth-interview consumer unit as a complete or incomplete reporter of income. The formal definition of "complete income reporter" status is fairly complex; Garner and Blanciforti (1994) give a detailed discussion. Current BLS procedure estimates mean income with the unadjusted mean response $\hat{\bar{Y}}_1$ defined by (1.1), with the R_i equal to indicators of complete income reporting, Y_i equal to income, and weights λ_i as described in Section 2.1. The weighted mean $\hat{\bar{Y}}_1$ uses both second- and fifth-interview data from a specified time period, but does not make direct use of the CE panel-data structure. In parallel with this, the present paper will distinguish between second- and fifth-interview data only in the construction of $\hat{\eta}_i$ and $\hat{Y}_i$ models.

Here, we used data from the second and fifth interview reports from all consumer units that had a second interview scheduled during 1990. The second-interview data involved 5,125 interviewed units and the fifth-interview data involved 5,093 interviewed units. For each interviewed unit (both the complete and the incomplete income reporters), BLS records provided a large number of demographic and expenditure variables; these were used as auxiliary variables in the modeling work described in Sections 3 and 4 below. For both the second and the fifth interviews, approximately 14 percent of the interviewed consumer units were incomplete income reporters.

3. CELLS BASED ON ESTIMATED RESPONSE PROBABILITIES

We first considered construction of adjustment cells based on estimated response probabilities. Logistic regression models for the complete-income-reporter probabilities $\eta_i = \eta(X_i)$ were fit separately for the second and fifth interview data described in Section 2. Model fitting details, including model parameter estimates and standard errors, are reported in Yansaneh and Eltinge (1993). All variance estimates were computed by the pseudo-replication method described in Section 2. The final model fits were used to estimate complete-reporter probabilities $\hat{\eta}_i$ for each second- and fifth-interview unit. Following the strategy in Section 1.3, units were grouped according to their $\hat{\eta}_i$ values into a total of k cells, with cell boundaries defined by the equal-quantile method.

3.1 Initial Sensitivity Analysis for the Number of Cells Used

The first three columns of Table 1 report the adjusted point estimates $\hat{\bar{Y}}_k$ of mean income, and associated standard errors, for several values of k. Comparisons of these point estimates indicate the extent to which the adjusted estimates are sensitive to a specific choice of k. For $k \geq 5$, the reported point estimates are relatively stable, varying between \$32,630 and \$32,664. This is consistent with the suggestion in Section 1.3 that $k = 5$ cells may provide most of the effective bias reduction to be obtained from a given cell-formation method; see Rosenbaum and Rubin (1984, Section 1 and Appendix A) for some related mathematical background.

In addition, note that for $k \geq 3$, the standard errors of $\hat{\bar{Y}}_k$ are also relatively stable, ranging from \$508 to \$530. This is in partial contrast with the general idea that selection of an

appropriate number of cells hinges on a bias-variance trade-off. For the present dataset, it appears that the effective bias reduction occurs fairly quickly (at $k = 5$, say), while substantial variance inflation does not occur until some point beyond $k = 20$. This is not unreasonable, since even for $k = 20$, the number of income responses per cell remained fairly large (ranging from 461 to 569), and thus avoided the general unstable-estimator problem associated with increasing numbers of sparse cells. Conversely, bias-variance tradeoff problems may be more severe for moderate k in applications involving smaller effective sample sizes, *e.g.*, estimation for small subpopulations.

Table 1
Adjusted Estimates of Mean Income with Cell Boundaries Determined by Estimated Response Probability Quantiles

Number of Cells	Point Estimate	Standard Error	$SE(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	MSE Ratio ($\hat{\gamma}_k$)
Unadjusted ($k = 1$)	32,967	569	N/A	N/A
$k = 3$ cells	32,736	530	112	1.30
$k = 4$ cells	32,779	518	122	1.28
$k = 5$ cells	32,630	523	138	1.53
$k = 6$ cells	32,664	515	122	1.51
$k = 10$ cells	32,640	514	116	1.58
$k = 15$ cells	32,638	515	118	1.58
$k = 20$ cells	32,634	508	118	1.63

3.2 Two Simple Cell Diagnostics

To complement the preceding sensitivity analysis, it is useful to study some sets of adjustment cells in additional detail. Let $C_1 = \{s_1,...,s_k\}$ be a given candidate set of adjustment cells, *e.g.*, the $k = 3$ or $k = 5$ equal-quantile- division cells in Section 3.1. The cells in C_1 can be refined by using equal-quantile divisions with a larger value of k; or by directly splitting one or more of the cells in C_1. This refinement may be worthwhile if there are empirical indications: (1) that the within-cell mean estimator $\bar{Y}_{hR}$ may be substantially biased; or (2) that a cell is wide relative to the precision with which the η_i values are estimated. Subsections 3.2.1 and 3.2.2 use two simple diagnostic methods to address issues (1) and (2), respectively. In each subsection, the proposed diagnostics lead to identification of potential "problem cells", and to construction of a refined set of adjustment cells, C_2, say. Comparisons of estimates of $\bar{Y}$ based on C_1 and C_2 then lead to some conclusions regarding the preferred set of $\hat{\eta}_i$-based adjustment cells.

3.2.1 Assessment of Within-Cell Bias

As noted in Section 1.2, a given adjusted estimator $\hat{\bar{Y}}_k$ reduces, but may not entirely eliminate, nonresponse bias; and the residual bias of $\hat{\bar{Y}}_k$ depends on the biases of the within-cell mean estimates $\bar{Y}_{hR}$. Consider the alternative within-cell mean estimator

$$\bar{Y}_{h\eta} = \left(\sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i\right)^{-1} \sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i Y_i. \tag{3.1}$$

If the $\hat{\eta}_i$ estimates were equal to the true response probabilities η_i, then (3.1) would be an approximately unbiased estimator of the true subpopulation mean $\bar{Y}_h$. In that case, an estimator of the within-cell bias $E(\bar{Y}_{hR} - \bar{Y}_h)$ would be $\hat{B}_h = \bar{Y}_{hR} - \bar{Y}_{h\eta}$, and the corresponding estimator of the overall bias $E(\hat{\bar{Y}}_k - \bar{Y})$ would be $\hat{B} = (\sum_{h=1}^{k} \sum_{j \in s_h} \lambda_j)^{-1} \sum_{h=1}^{k} (\sum_{j \in s_h} \lambda_j) \hat{B}_h$.

Because the $\hat{\eta}_i$ values are subject to estimation error, the terms $\hat{B}_h$ and $\hat{B}$ give only a partial indication of potential bias problems. For example, a large value of $\hat{B}_h$ may reflect a substantial bias in $\hat{\bar{Y}}_{hR}$, or may reflect biases in the alternative estimator $\bar{Y}_{h\eta}$ due to the errors $\hat{\eta}_i - \eta_i$; *cf.* the cautionary remarks in Little (1986, p. 146) regarding direct use of the weights $\hat{\eta}_i^{-1}$ in adjusted estimation of $\bar{Y}$. Thus, if one observes a large value of $\hat{B}_h$, it is worthwhile to consider refinement of cell h; but the final decision of whether to use the resulting refined set of cells will depend on whether the refined set produces a substantially different estimate of the overall mean $\bar{Y}$.

Tables 2 and 3 present $\hat{B}_h$ values and associated standard errors and t statistics for equal-quantile-division cells with $k = 3$ and $k = 5$, respectively. Note that for the case $k = 3$, the $\hat{B}_h$ diagnostics indicate a possible bias contribution from the lowest cell. This is consistent with the suggestion from Section 3.1 that $k = 3$ cells may not provide a satisfactory nonresponse adjustment. In addition, the corresponding value of $\hat{B}$ was 111, with a standard error of 75; this value of $\hat{B}$ is very close to the difference $\hat{\bar{Y}}_3 - \hat{\bar{Y}}_5 = 106$ of the estimates $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_5$ from Table 1.

Table 2
Within-Cell $\hat{B}_h$ Statistics for Probability-Based Cells, $k = 3$

h	$\hat{B}_h$	$se(\hat{B}_h)$	$t = \hat{B}_h/se(\hat{B}_h)$
1	269	136	1.98
2	-19	43	-0.44
3	84	45	1.87

Table 3
Within-Cell $\hat{B}_h$ Statistics for Probability-Based Cells, $k = 5$

h	$\hat{B}_h$	$se(\hat{B}_h)$	$t = \hat{B}_h/se(\hat{B}_h)$
1	96	217	0.44
2	-72	116	-0.62
3	-52	56	-0.93
4	-16	27	-0.59
5	98	50	1.96

In light of the preceding results, the low-$\hat{\eta}_i$ cell from the $k = 3$ case was split in half. The upper bounds for the two new cells ($h = 1'$ and $h = 1''$, say) were determined by the

0.167 and 0.333 estimated quantiles of the $\hat{\eta}_i$ population. The resulting $\hat{B}_h$ values and standard errors were 90 and 197 for cell 1′, and −42 and 79 for cell 1″. In addition, the refined set of four cells had $\hat{B} = 30$, with a standard error of 75; and the adjusted estimate of $\bar{Y}$ equal to \$32,652 and standard error of \$518 were close to those obtained from the equal-quantile-division method with $k = 5$.

In contrast with the results for $k = 3$, the $\hat{B}_h$ results for $k = 5$ indicated relatively little basis for concern, with the possible exception of cell $h = 5$, which had a t statistic of 1.96. For $k = 5$, the value of $\hat{B}$ was 11, with a standard error of 93. Additional splitting of cell $h = 5$ did not lead to notable changes in either the estimate of $\bar{Y}$ or the associated standard errors. The $\hat{B}_h$ results, for equal-quantile-division cells with larger values of k showed even fewer indications of within-cell bias. For example, for $k = 6$ all six cells had $\hat{B}_h$ values with t statistics less than or equal to 1.65; and for $k = 10$, all cells had $\hat{B}_h$ values with t statistics less than or equal to 1.54.

3.2.2 Relation of Cell Widths to Precision of η_i Estimates

The relationship between the widths of adjustment cells and the widths of confidence intervals for the response probabilities η_i leads to another diagnostic for identification of potential problem cells. First, define $a_h = (\sum_{i\in s_h}\lambda_i R_i)^{-1}\sum_{i\in s_h}\lambda_i$, the nonresponse-adjustment factor used for responding units in cell h. Second, following standard results for logistic regression, note that an approximate 95% confidence interval for η_i is

$$(LB_i, UB_i) = ([1+\exp\{-X_i'\hat{\theta}+1.96D_i^{1/2}\}]^{-1}$$

$$[1+\exp\{-X_i'\hat{\theta}-1.96D_i^{1/2}\}]^{-1}),$$

where $\hat{\theta}$ is the vector of logistic regression parameter estimates, $D_i = X_i'\hat{V}_\theta X_i$, and $\hat{V}_\theta$ is the pseudo-replicate-based estimated covariance matrix for $\hat{\theta}$. Let $\bar{d}_h$ be the λ_i-weighted sample mean of the confidence interval widths $UB_i - LB_i$ for units i in cell h, and consider a comparison of $\bar{d}_h$ to the width of cell h. If cell h is relatively wide, both on an absolute scale and relative to $\bar{d}_h$, then division of this cell may produce two new cells with two substantially different weight factors a_h. Conversely, if $\bar{d}_h$ is substantially larger than the width of cell h, then differences among $\hat{\eta}_i$ in that cell may result more from estimation error than from differences in the true η_i. In that case, additional division of cell h is unlikely to produce much useful change in weight factors a_h; and thus there will be relatively little change in the resulting nonresponse-adjusted estimator of $\bar{Y}$.

Tables 4 and 5 report cell boundaries, cell widths, $\bar{d}_h$, and a_h values for $k = 5$ and $k = 10$, respectively. For $k = 5$, the widths of cells 2 through 5 were not large relative to the $\bar{d}_h$ values. Each of these cells is essentially split in half to produce the $k = 10$ cell case. The resulting pairs of a_h for $k = 10$ are relatively close to the corresponding a_h values in cells 2 through 5 with $k = 5$.

By contrast, with $k = 5$, cell 1 is over twice as wide as $\bar{d}_1$. When $k = 10$, this cell is divided into cells with somewhat different nonresponse adjustment weight factors a_h: 1.45 and 1.27, respectively. However, the corresponding cell-mean estimates are relatively close: $\bar{Y}_{1R} = \$24{,}045$ and $\bar{Y}_{2R} = \$24{,}582$ for $k = 10$. Thus, in this example, the nonresponse-adjusted estimates $\hat{\bar{Y}}_5$ and $\hat{\bar{Y}}_{10}$ are relatively close because four of the five cell divisions produced relatively small changes in weights, and because the other cell division produced two cells with similar cell means.

Table 4
Estimated-Probability Cell Boundaries, Cell Widths, Mean Confidence Interval Widths and Nonresponse Adjustment Factors, $k = 5$

h	Lower Bound	Upper bound	Cell Width	$\bar{d}_h$	a_h
1	0.384	0.810	0.426	0.197	1.35
2	0.810	0.861	0.051	0.139	1.20
3	0.861	0.894	0.033	0.110	1.13
4	0.894	0.924	0.030	0.088	1.08
5	0.924	0.994	0.070	0.067	1.07

Finally, the a_h factors in Table 5 indicate that mean response rates in the $k = 10$ cells fall in a moderate range, from $(1.45)^{-1} = 0.69$ to $(1.06)^{-1} = 0.94$. Some other nonresponse datasets involve a wider range, and thus are more likely to produce more pronounced cell-splitting results. Conversely, other nonresponse datasets may display a tighter distribution of response probabilities, and thus are less likely to display notable cell-splitting effects.

Table 5
Estimated-Probability Cell Boundaries, Cell Widths, Mean Confidence Interval Widths and Nonresponse Adjustment Factors, $k = 10$

h	Lower Bound	Upper Bound	Cell Width	$\bar{d}_h$	a_h
1	0.384	0.762	0.378	0.220	1.45
2	0.762	0.810	0.048	0.174	1.27
3	0.810	0.840	0.030	0.146	1.21
4	0.840	0.861	0.021	0.132	1.19
5	0.861	0.878	0.017	0.111	1.14
6	0.878	0.894	0.016	0.108	1.11
7	0.894	0.908	0.014	0.093	1.09
8	0.908	0.924	0.016	0.083	1.08
9	0.924	0.944	0.020	0.072	1.08
10	0.944	0.994	0.050	0.062	1.06

3.3 Comparison of Cell-Based Estimates to the Unadjusted Estimate

To conclude the assessment of $\hat{\eta}_i$-based cells, we compared the adjusted estimates $\hat{\bar{Y}}_k$ with the unadjusted

estimate $\hat{\bar{Y}}_1$. First, Table 1 indicates that for the reported values of $k \geq 5$, the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are greater than or equal to \$303. Second, for $k \geq 5$, the estimated standard errors of the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are all less than or equal to \$138, and the corresponding t statistics are all greater than 2.44. Thus, for $k = 5$, say, a formal test of the hypothesis H_0: $E(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_5) = 0$ would be rejected at standard significance levels; *i.e.*, the adjustment-cell method has produced a significant change in the mean income estimate.

In addition, a rough comparison of the efficiencies of $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_k$ follows from the estimated mean squared error ratio

$$\hat{\gamma}_k = \{\hat{V}(\hat{\bar{Y}}_k)\}^{-1}[\hat{V}(\hat{\bar{Y}}_1) + \max\{0, (\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)^2 - \hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)\}]$$

where $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, and $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ are the pseudo-replicate-based variance estimates for the indicated means. To interpret this ratio, assume for the moment that $\hat{\bar{Y}}_k$ is an approximately unbiased estimator of $\bar{Y}$. Then $\hat{\gamma}_k$ is an estimator of the mean squared error of the unadjusted estimator $\hat{\bar{Y}}_1$, relative to the mean squared error of $\hat{\bar{Y}}_k$. Consequently, $\hat{\gamma}_k$ reflects the loss of efficiency incurred by using the biased, unadjusted estimator $\hat{Y}_1$ instead of the adjusted, unbiased estimator $\hat{\bar{Y}}_k$. However, this interpretation should be viewed with some caution, since it depends on the assumption that $\hat{\bar{Y}}_k$ is approximately unbiased for $\bar{Y}$, and since the $\hat{\gamma}_k$ are functions of the random terms $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$, $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, and $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$.

As suggested by a referee, one could also consider a mean squared error ratio

$$\{\hat{V}(\hat{\bar{Y}}_\eta)\}^{-1}[\hat{V}(\hat{\bar{Y}}_k) + \max\{0, (\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)^2 - \hat{V}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)\}]$$

where $\hat{\bar{Y}}_\eta$ equals expression (1.1) with λ_i replaced by $(\hat{\eta}_i)^{-1}\lambda_i$. This would amount to comparing each cell-based estimate $\hat{\bar{Y}}_k$ to $\hat{\bar{Y}}_\eta$. This is appropriate if $\hat{\bar{Y}}_\eta$ is approximately unbiased, but this unbiasedness may be problematic in some cases; *cf.* Little (1986, p. 146).

The final column of Table 1 reports the estimated ratios $\hat{\gamma}_k$ for specified values of k. For $k \geq 5$, each reported $\hat{\gamma}_k$ is greater than 1.5. Finally, note that each adjusted estimate $\hat{\bar{Y}}_k$ fell *below* the unadjusted estimate $\hat{\bar{Y}}_1$. This occurred because, for a given k, cells associated with larger response probabilities tended to have larger mean estimates $\bar{Y}_{hR}$. For example, for $k = 5$, the $\bar{Y}_{hR}$ values were \$24,333, \$33,729, \$33,398, \$34,620, and \$37,057 for $h = 1$ (the low $\hat{\eta}_i$ cell) through $h = 5$ (the high $\hat{\eta}_i$ cell), respectively.

4. CELLS BASED ON ESTIMATED INCOME VALUES

The general diagnostic ideas of Section 3 also apply to $\hat{Y}_i$ based cells. To illustrate this idea, we fit separate weighted regressions of Y_i = reported income for second- and fifth-interview respondents. Yansaneh and Eltinge (1993) report details of the work, including parameter estimates and standard errors. The resulting regression models were used to compute estimated incomes $\hat{Y}_i$ for both complete and incomplete income reporters. Units were then grouped into cells according to their $\hat{Y}_i$ values, with cell boundaries determined by the equal-quantile method.

Table 6 reports the basic sensitivity-analysis and efficiency results for the $\hat{Y}_i$ based cells; the organization of this table is the same as in Table 1. The sensitivity-analysis results are qualitatively similar, but not identical, to those reported for the $\hat{\eta}_i$-based cells. In additional work not detailed here, we considered splitting individual equal-quantile $\hat{Y}_i$-based cells. For $k \geq 4$, the resulting mean estimates and associated standard errors did not differ notably from those reported in Table 6.

Table 6

Adjusted Estimates of Mean Income with Cell Boundaries Determined by Estimated Income Quantiles

Adjustment Method	Point Estimate	Standard Error	$SE(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	MSE Ratio
Unadjusted ($k = 1$)	32,967	569	N/A	N/A
$k = 3$ cells	32,512	509	106	2.01
$k = 4$ cells	32,468	512	108	2.14
$k = 5$ cells	32,473	511	115	2.12
$k = 6$ cells	32,492	508	117	2.08
$k = 10$ cells	32,488	510	119	2.07
$k = 15$ cells	32,478	504	124	2.16
$k = 20$ cells	32,495	513	124	2.02

The final two columns of Table 6 permit comparison of $\hat{\bar{Y}}_k$ to the unadjusted estimate $\hat{\bar{Y}}_1$. For $k \geq 4$, the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ are greater than or equal to \$472, with estimated standard errors less than or equal to \$124. The associated t statistics are all greater than 3.80. In addition, the estimated mean squared error ratios $\hat{\gamma}_k$ are all greater than 2.0.

Also, the $\hat{\eta}_i$ and $\hat{Y}_i$-based cells produced somewhat different adjusted estimates of mean income, but the observed differences were not statistically significant at customary α levels. For example, with $k = 5$, the difference between the $\hat{\eta}_i$- and $\hat{Y}_i$-based cell estimates is \$32,630 - \$32,473 = \$157, with a standard error of \$122 and a t statistic of 1.29. Similarly, for $k = 10$, the difference between the $\hat{\eta}_i$- and $\hat{Y}_i$-based estimates is \$152, with a standard error of \$104. Thus, the data provide relatively little power to distinguish between results of the two general cell-formation methods.

Finally, note that a given set of $\hat{Y}_i$-based cells are fundamentally linked with a particular Y variable, *e.g.*, consumer unit income. Consequently, that set of cells will not necessarily work well for estimation of the mean of a different Y variable.

5. DISCUSSION

5.1 Summary of Methods

This paper has discussed some simple diagnostics for formation of nonresponse adjustment cells. The methodology may be summarized as follows.

1. Based on preliminary modeling work and observed auxiliary variables X_i, compute an estimated response probability $\hat{\eta}_i$ for each sample unit (respondents and nonrespondents).
2. Construct k adjustment cells with boundaries determined by the estimated $k^{-1}j$ quantiles of the $\hat{\eta}_i$ population, $j = 1, 2, \ldots, k - 1$. Compute the resulting adjusted mean estimate, $\hat{\bar{Y}}_k$.
3. Repeat (2) for several integers $k > 1$. As k increases, identify the point at which the $\hat{\bar{Y}}_k$ become approximately constant. In keeping with Rosenbaum and Rubin (1984) and the empirical results discussed here, values of k near 5 may be of special interest.
4. Use simple screening diagnostics (*e.g.*, $\hat{B}_h$ and $\bar{d}_h$ in Section 3.2) to check for potential problems in the equal-quantile-division adjustment cells. If the diagnostics identify potential "problem cells," then try additional refinements of these cells. Compute estimates of $\bar{Y}$ based on these refined sets of cells, and compare these new estimates to the $\hat{\bar{Y}}_k$ from (3).
5. Assess the overall effect of adjustment by comparing the differences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ to the standard errors $se(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$; and by computing the estimated mean squared error ratios $\hat{\gamma}_k$.
6. Repeat steps (1) through (5), as appropriate, for $\hat{Y}_i$-based adjustment cells. Compare the final estimates of $\bar{Y}$ obtained from the $\hat{\eta}_i$ and $\hat{Y}_i$-based cell methods.

5.2 Areas for Future Research

The results of this work suggest two potentially useful areas for future research. First, the CE income nonresponse problem is similar to nonresponse problems in some other large-scale surveys, but as with any case study one should not over-generalize the empirical results reported here. It would be useful to apply these diagnostics to problems involving different estimands (*e.g.*, cross-class means) or involving nonresponse datasets with somewhat different characteristics, *e.g.*, larger or smaller effective sample sizes; or wider or narrower distributions of $\hat{\eta}_i$ estimates. This in turn would offer additional insight into the operating characteristics of $\hat{\eta}_i$ and $\hat{Y}_i$-based adjustment cell methods in practical applications. Second, extensions to multivariate problems (*e.g.*, relationships involving second-interview and fifth-interview CE income data) also would be of interest.

ACKNOWLEDGEMENTS

The authors thank Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott, and Stephanie Shipp for many helpful discussions of the Consumer Expenditure Survey; and Wayne Fuller, Steve Miller, Geoff Paulin, Stuart Scott, three referees and the editor for helpful comments on earlier versions of this paper. This work was carried out while the authors were visiting the Bureau of Labor Statistics through the ASA/NSF/BLS Research Fellow Program, and was supported by a grant from the National Science Foundation (SES-9022443). Eltinge's research was also supported in part by a grant from the National Institutes of Health (CA 57030-04). The views expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of Labor Statistics.

REFERENCES

CASSEL, C.-M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys*, (Vol. 3), (Eds. W.G. Madow, I. Olkin, and D. Rubin). New York: Academic Press, 143-160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205-213.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R.J.A., SAMUHEL, M., and TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 168-173.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

EZZATI, T., and KHARE, M. (1992). Nonresponse adjustments in a national health survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 339-344.

GARNER, T.I., and BLANCIFORTI, L.A. (1994). Household income reporting: An analysis of U.S. Consumer Expenditure Survey data. *Journal of Official Statistics* 10, 69-91.

GÖKSEL, H., JUDKINS, D.R., and MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

KALTON, G., and MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 409-428.

LEPKOWSKI, J., KALTON, G., and KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P.R., and RUBIN, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

TREMBLAY, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology*, 12, 85-97.

UNITED STATES BUREAU OF LABOR STATISTICS (1991). News: Consumer Expenditures in 1990. Publication USDL 91-607, United States Department of Labor, Washington, DC.

UNITED STATES BUREAU OF LABOR STATISTICS (1992). BLS Handbook of Methods. Bulletin 2414, United States Department of Labor, Washington, DC.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YANSANEH, I.S., and ELTINGE, J.L. (1993). Construction of adjustment cells based on surrogate items or estimated response propensities. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 538-543.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 41-52
Statistics Canada

Variance Estimation for Measures of Income Inequality and Polarization – An Empirical Study

MILORAD S. KOVAČEVIĆ and WESLEY YUNG[1]

ABSTRACT

Measures of income inequality and polarization are fundamental to the discussions of many economic and social issues. Most of these measures are non-linear functions of the distribution function and/or the quantiles and thus their variances are not expressible by simple formulae and one must rely on approximate variance estimation techniques. In this paper, several methods of variance estimation for six particular income inequality and polarization measures are summarized and their performance is investigated empirically through a simulation study based on the Canadian Survey of Consumer Finance. Our findings indicate that for the measures studied here, the bootstrap and the estimating equations approach perform considerably better than the other methods.

KEY WORDS: Gini index; Lorenz curve ordinate; Low income proportion; Polarization index; Quantile share; Resampling variance estimation; Linearization method.

1. INTRODUCTION

Analyses of the distribution of income are fundamental to the discussions of important economic and social issues such as the extent of inequality, poverty, the size of the middle class, *etc*. There exists extensive statistical and econometric literature on this subject, especially on different measures of income inequality and their properties (Sen 1973, Kakwani 1980, Nygärd and Sandström 1981). However, seldom is there any attempt to produce information regarding the sampling variability associated with the estimates used to assess the magnitude of inequality or polarization. Such information is necessary for two reasons: i) as a measure of the precision of the estimates obtained from survey data and ii) to provide a basis for formal statistical inference on income distributions, particularly when income distributions are compared over different regions or across time.

Measures of income inequality and polarization are finite population parameters expressible as functions of the ordered population values, thus their variances are not obtainable in simple formulae and one has to rely on approximate variance estimation techniques. Generally, inference about these measures, based on a complex sample design, embodies point estimation and confidence intervals. We investigate variance estimation for some of these measures such as quantiles, low income line, low income proportion, Lorenz curve ordinates, quantile shares, Gini index, and the polarization index.

Throughout this paper we assume a fixed finite population framework, that is, we assume that associated with each population unit is a fixed but unknown real number: the value of income earned by the unit. We assume that the population is stratified into L strata with N_h primary sampling units (PSU's) in the h-th stratum. In the first stage sample, $n_h (\geq 2)$ PSU's are selected from stratum h (independently across strata). We assume that subsampling within sampled PSU's is performed to ensure unbiased estimation of PSU totals, $Y_{hc}, c = 1,...,n_h; h = 1,...,L$. Attached to the ($hci$)-th ultimate unit, along with the observed variable of interest, y_{hci}, is the sampling weight w_{hci}. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample, incorporating all stages of sampling.

After reviewing the basic definitions of these measures, we give their point estimates under our sample design in section 2. Section 3 deals with variance estimation of these measures. Existing methods are reviewed and five methods, jackknifing, grouped and repeatedly grouped balanced half-sample, bootstrap and linearization via the estimating equations approach are summarized in detail. Section 4 contains the description of the simulation study based on data collected in the 1988 Canadian Survey of Consumer Finance. The empirical study is aimed at comparisons of the variance estimation methods for a number of income inequality measures. Various results are presented, summarized and interpreted. Our conclusions are presented in section 5.

2. ESTIMATION OF INCOME INEQUALITY MEASURES

The simplest measures of inequality between two distributions are the cumulative distribution function (CDF) and the quantiles of the two distributions. We start this section by defining the CDF and the finite population quantiles. The remaining measures studied in this paper are functions of the CDF or a fixed number of quantiles and are introduced in section 2.1.

For a variable Y defined over a finite population $U = \{1,...,N\}$, we define the CDF as

[1] Milorad S. Kovačević, Senior Methodologist, Household Survey Methods Division, and Wesley Yung, Senior Methodologist, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

$$F_N(y) = \sum_{i \in U} I\{Y_i \le y\} \frac{1}{N},$$

where $I\{a\}$ is an indicator function taking on a value of 1 if a is true and 0 otherwise. A design unbiased estimator of $F_N(y)$ is

$$\tilde{F}(y) = \sum_{i \in s} I\{y_i \le y\} \frac{w_i}{N}$$

where the sampling weights, w_i, are obtained from the sample design and are equal to the inverse of the first order inclusion probabilities. This estimator may not be a CDF since $\tilde{F}(\infty) = \hat{N}/N$ may not necessarily be equal to 1. Thus we would rather use the possibly design-biased estimator:

$$\hat{F}(y) = \sum_{i \in s} I\{y_i \le y\} w_i \Big/ \sum_{i \in s} w_i = \sum_{i \in s} I\{y_i \le y\} \tilde{w}_i, \qquad (2.1)$$

where $\tilde{w}_i = w_i / \sum_s w_i$, $i \in s$. The estimator (2.1) is design unbiased when $\sum_s w_i = N$ which can occur under simple random sampling or if the weights, w_i, are benchmarked to known population totals. In general, the estimator (2.1) uses final weights which usually involve poststratification, non-response adjustment, some iterative calibrations and so on. In this paper, we consider only the case where the design weights are used.

Turning to the quantiles, we define the finite population quantiles as

$$\xi_N(p) = \inf_{i \in U} \{Y_i \mid F_i \ge p\} \quad \text{for} \quad 0 < p \le 1,$$

where $F_i = F_N(Y_i)$. The population quantiles are estimated by the sample quantiles

$$\hat{\xi}_p = \inf_{i \in s} \{y_i \mid \hat{F}_i \ge p\} \quad \text{for} \quad 0 < p \le 1,$$

where $\hat{F}_i = \hat{F}(y_i)$. If a parameter is a function of quantiles, say $\theta_N = g(\underset{\sim}{\xi}_N)$ with $\underset{\sim}{\xi}_N = \{\xi_N(p_1), ..., \xi_N(p_k)\}$, then it is estimated by $\hat{\theta} = g(\underset{\sim}{\hat{\xi}})$ where $\underset{\sim}{\hat{\xi}} = (\hat{\xi}_{p_1}, ..., \hat{\xi}_{p_k})$.

2.1 Income Inequality and Polarization Measures as Finite Population Parameters

In this section we present some frequently used income inequality and polarization measures. They are the low income line, the low income proportion, the Lorenz Curve and its related statistics, the quantile shares, the Gini index and finally the polarization curve and the polarization index. Our intention is to briefly introduce these measures, not to discuss them in detail. For more details, we refer the readers to Nygård and Sandström (1981) and Wolfson (1994).

The *low income line*, or the *poverty line*, is defined as a fraction of the median, $\lambda_\alpha = \alpha\, \xi_N(0.5)$, where $0 < \alpha \le 1$ is a given constant and $\xi_N(0.5)$ is the finite population median. Its estimate is simply $\hat{\lambda}_\alpha = \alpha\, \hat{\xi}_{0.5}$.

The *low income proportion (LIP)* is the percentage of units (individuals, families, households) in the population falling below the low income line λ_α and is given by $\Lambda_\alpha = F_N(\lambda_\alpha)$. The estimate of the low income proportion involves the estimation of both the distribution function and the low income line, $\hat{\Lambda}_\alpha = \hat{F}(\hat{\lambda}_\alpha) = \sum_s I\{y_{hci} \le \alpha\, \hat{\xi}_{0.5}\} \tilde{w}_{hci}$.

The finite population *Lorenz curve ordinate (LCO)* gives the share of income received by the poorest $100p$ percent of the population and is defined as a function of p $(0 \le p \le 1)$. It simply depicts the cumulative income against the population share. As a parameter it is defined as

$$L(p) = \frac{1}{\mu_Y} \int_0^p \xi_q \, dq$$

where μ_Y is the population mean, and ξ_q is the quantile function. For a large population without ties the expression above is approximated by

$$L_N(p) \approx \sum\nolimits_U \frac{I\{F_i \le p\}\, Y_i}{\mu_N} \frac{1}{N}$$

and estimated as

$$\hat{L}(p) = \sum\nolimits_s \frac{I\{\hat{F}_{hci} \le p\}\, y_{hci}}{\hat{\mu}} \tilde{w}_{hci}$$

where $\hat{\mu} = \sum_s \tilde{w}_{hci} y_{hci}$ and $\hat{F}_{hci} = \hat{F}(y_{hci})$.

The *quantile share (QS)* is defined as the proportion of total income shared by the population allocated to a quantile interval $[\xi_{p_1}, \xi_{p_2})$:

$$Q_N(p_1,p_2) \approx \sum\nolimits_U \frac{I\{ep_1 \le F_i \le p_2\}\, Y_i}{\mu_N} \frac{1}{N} = L_N(p_2) - L_N(p_1)$$

For $0 \le p_1 < p_2 \le 1$ it is estimated by replacing the parameters with their estimates.

The most popular measure of aggregate inequality of income distribution, the *Gini index*, is defined as the area between the Lorenz curve and the 45° line, normalized to lie between 0 and 1: $G = 1 - 2\int_0^1 L(p)\,dp$. Its finite population version is estimated by

$$\hat{G} = \sum\nolimits_s \frac{[2\hat{F}_{hci} - 1]\, y_{hci}}{\hat{\mu}} \tilde{w}_{hci}.$$

For more about the Gini index we refer the reader to Nygård and Sandström (1985).

Using the analogy of the Lorenz curve and the Gini index, Foster and Wolfson (1992) defined the *polarization curve* as

$$B(p) = \int_{0.5}^{p} \frac{F^{-1}(q) - \xi_{0.5}}{\xi_{0.5}} \, dq,$$

or in the finite population form

$$B(p) = \begin{cases} 0.5 - p - \dfrac{1}{\xi_{0.5}} \sum\limits_U I\{p < F_i < 0.5\}\, Y_i \dfrac{1}{N}, & 0 < p \le 0.5, \\ 0.5 - p + \dfrac{1}{\xi_{0.5}} \sum\limits_U I\{0.5 \le F_i < p\}\, Y_i \dfrac{1}{N}, & 0.5 < p \le 1. \end{cases}$$

The polarization curve shows, for any population percentile, how far its income is from the median. The area below the polarization curve is considered as a summary measure of the polarization. A version of it, normalized to lie between 0 and 1, is named the *polarization index* (PI):

$$PI_N = \sum_U \frac{[2 - 2I\{F_i \le 0.5\} - 2F_i]Y_i}{\xi_N(0.5)} \frac{1}{N}$$

where $\xi_N(0.5), \mu_N$ and F_i were previously defined. The estimate of the polarization index is obtained by replacing the parameters with their estimates.

3. VARIANCE ESTIMATION

The estimation of the variance of non-smooth statistics like quantiles, as well as quantile based functions like the low income proportion or the polarization index, is not straightforward especially when the assumption of simple random sampling is untenable and there is a need to take into account the complex sample design. In the first part of this section we review some results on variance estimation for quantiles as a starting point for understanding the complexity of variance estimation for income inequality measures. We also review results on variance estimation for some measures like the Lorenz curve ordinates. The second part describes the methods of variance estimation that are used in this study.

Woodruff (1952) proposed a method to obtain confidence intervals for individual quantiles. These intervals were used by Francisco and Fuller (1986) and Rao and Wu (1987) to derive variance estimators. Though the estimator depends on the confidence coefficient, Rao and Wu (1987) established its asymptotic consistency for any significance level α. Using Monte Carlo simulations, they studied the standard errors of quantiles for cluster samples estimated in this manner. Their results suggest that a 95% confidence interval works well as a basis for extracting the standard error. Binder (1991) obtained a similar form of the variance estimator by using the linearization method.

Jackknife variance estimators have become extremely popular for smooth functions of totals and means with the increase in computing power. Standard asymptotic theory applied to the median of a distribution with bounded continuous density, f, shows that $nE(\hat{\xi}_{0.5} - \xi_{0.5})^2 \to 1/[4f^2(\xi_{0.5})]$ as $n \to \infty$. Efron (1979) pointed out that the jackknife method applied to the sample median gives a variance estimate which is asymptotically inconsistent since

$$n\,\mathrm{var}_{JK}(\hat{\xi}_{0.5}) \to \frac{1}{4f^2(\xi_{0.5})}[\chi_2^2/2]^2$$

where $[\chi_2^2/2]^2$ has mean of 2 and variance of 20 which means that the jackknife variance estimator tends to over estimate, on the average, the correct asymptotic variance by 100%. Kovar (1987) confirmed empirically the inconsistency of the delete-one-unit jackknife estimators for a stratified sample design. In a simulation study using a stratified population, he showed that the delete-one-unit jackknife estimators (he considered six of them) performed poorly, over estimating the true variance by 30-70% in the design with two units per stratum and performed even worse in the five units per stratum design. Shao and Wu (1989), however, have shown that under certain conditions, the delete-d jackknife method has desirable asymptotic properties for variance estimation of non-smooth statistics. This result has motivated Rao, Wu and Yue (1992) to apply the delete-one-PSU jackknife for stratified multistage sampling. In a limited simulation study they found that both bias and relative bias of the jackknife variance estimator of the median decrease as the cluster size increases for a fixed intracluster correlation.

Bootstrap variance estimation for the median was first reported by Efron (1979), and in the case of independent and identically distributed observations the bootstrap provides consistent results, (see also Babu 1986). Rao and Wu (1988) gave a modified bootstrap method for variance estimation in stratified designs. Kovar (1987) and Kovar, Rao and Wu (1988) reported good performance for medians when the size of the bootstrap sample is $n_h^* = n_h - 1$.

In the grouped balanced half-sample method (GBHS) of variance estimation, the sampled clusters in each stratum are randomly divided into two groups (halves) and the balanced repeated replication method is applied to the groups. Rao and Shao (1996) showed that this method is asymptotically incorrect in the sense that the associated t-pivotal does not converge in distribution to a standard normal distribution and that the associated confidence intervals are asymptotically incorrect. To overcome this difficulty they proposed independently repeating the grouping T times and then taking the average of the resulting T variance estimates. They showed the asymptotic correctness of such an estimator for a stratified random sampling design as $\min n_h \to \infty$ and $T \to \infty$. In a small simulation study they found that the method performs well for T as small as 15 in the case of smooth estimators. For a variance estimator of the population median, the RGBHS method performed better than the jackknife and GBHS in the sense that the RGBHS had a smaller relative bias and a smaller coefficient of variation. Recently, McCarthy (1993) discussed and compared a variety of procedures for variance estimation of the median based on simple random samples drawn from a finite population without replacement. His study includes most resampling procedures.

Although, the linearization methods useful for nonlinear statistics are difficult to implement for quantiles since density estimation is involved, Binder (1991), Binder and Kovačević (1995) and Kovačević and Binder (1997) obtained consistent estimators for the variance of some non-smooth measures of income inequality and polarization using the linearization method within the estimating equation framework. Estimators obtained using this method are computationally simpler than the resampling estimators but require theoretical derivation.

Variance estimation of the Gini Index has been studied by several authors under the assumption of simple random sampling, Glasser (1962), Sendler (1979), Sandström, Wretman and Waldén (1985) and Yitzhaki (1991). In the case of a complex design, Love and Wolfson (1976) proposed a 'crude half-sample replication' method. Sandström, Wretman and Waldén (1988) compared approximate variance techniques with the delete-one-unit jackknife for three sampling designs, two of which were complex.

Estimation of the variance of the Lorenz curve ordinates and the corresponding quantile shares has received less attention. The derivation of their asymptotic variances is quite complicated. There is the pioneering work of Beach and Davidson (1983) and Beach and Kaliski (1986). Their work is based on the superpopulation framework in which the survey weights are seen as constants in the construction of estimates. This approach, due to its model-based nature, may have its limitations in applications to data obtained from sample surveys where the sample design is deemed to be significant.

In the following subsections we review the variance estimation methods used in this study.

3.1 Delete-one-PSU Jackknife

This method is based on the sequential exclusion (deletion) of one PSU at a time from the computation of the estimate. After deletion, the weights of the remaining units in the sample are modified in such a manner that the deleted weights are compensated and that the CDF estimated from the remaining sample has the same properties of the original CDF. Let $\hat{F}_{(gj)}(y)$ denote the estimate of the CDF based on a sample without the gj-th PSU, that is

$$\hat{F}_{(gj)}(y) = \hat{G}_{(gj)}(y) / \hat{N}_{(gj)}$$

where

$$\hat{G}_{(gj)}(y) = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} I\{y_{hci} \leq y\} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci} I\{y_{gci} \leq y\}$$

and

$$\hat{N}_{(gj)} = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci}.$$

The 'delete-one-PSU' jackknife variance estimator of $\hat{F}(y)$ is

$$v_{J1}(\hat{F}(y)) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{F}_{(gj)}(y) - \hat{F}(y))^2.$$

Asymptotic consistency of $v_{J1}(\hat{F}(y))$ can be established using results from Krewski and Rao (1981).

For convenience, we note that all measures considered here can be written in the general form

$$\theta_N = \sum_{U} J(F_N, Y_i, \boldsymbol{\beta}) \frac{1}{N},$$

where $J(\cdot)$ is a real-valued function possibly dependent on the nuisance parameter, $\boldsymbol{\beta}$. The finite population parameter θ_N is then estimated by

$$\hat{\theta} = \sum\nolimits_s J(\hat{F}, y_{hci}, \hat{\boldsymbol{\beta}}) \tilde{w}_{hci} \tag{3.1}$$

where $\hat{\boldsymbol{\beta}}$ denotes the estimated vector of nuisance parameters and $\tilde{w}_{hci}$ are the standardized weights. Using this general form, the estimate of an income inequality measure computed from the sample after omitting PSU gj, is

$$\hat{\theta}_{(gj)} = \sum\nolimits_s J(\hat{F}_{(gj)}, y_{hci}, \hat{\boldsymbol{\beta}}_{(gj)}) \tilde{w}_{hci(gj)}$$

where $\hat{F}_{(gj)}$ and $\hat{\boldsymbol{\beta}}_{(gj)}$ are the values of the distribution function and the nuisance parameter estimated from the sample with the gj-th PSU deleted and

$$\tilde{w}_{hci(gj)} = \begin{cases} w_{hci}/\hat{N}_{(gj)}, & \text{if } h \neq g, \\ \frac{n_g}{n_g - 1} w_{gci}/\hat{N}_{(gj)}, & \text{if } h = g, c \neq j, \\ 0, & \text{if } h = g, c = j. \end{cases}$$

The resulting 'delete-one-PSU' jackknife variance estimator of $\hat{\theta}$ is

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2. \tag{3.2}$$

If $\hat{\theta}$ is substituted by $\hat{\theta}_{..} = \sum_g \sum_j \hat{\theta}_{(gj)}/n$ a variant of the jackknife variance estimate is obtained. We denote it by $v_{J2}(\hat{\theta})$. Obviously $v_{J2}(\hat{\theta}) \leq v_{J1}(\hat{\theta})$. The consistency of (3.2) for smooth statistics has been established by Krewski and Rao (1981).

In the case of variance estimation for quantiles and functions of quantiles, we first compute the quantiles based on the sample with the gj-th PSU deleted,

$$\hat{\xi}_{(gj)}(p) = \inf\{y_{hci} \mid \hat{F}_{(gj)}(y_{hci}) \geq p, hci \in s \setminus (gj)\},$$

compute $\hat{\theta}_{(gj)} = g(\hat{\xi}_{(gj)})$ and then use equation (3.2) to obtain a jackknife variance estimator.

3.2 Grouped Balanced Half-Sample (GBHS) Method and Repeatedly Grouped Balanced Half-Sample (RGBHS) Method

Originally, the balanced half-sample method was proposed for the two clusters-per-stratum designs. The case that we are interested in is when there are more than 2 clusters per stratum. This situation is usually handled by grouping the clusters (primary stage units) in each stratum into two groups. We explore the idea given by Wu (1991) and simplify its application for the variance estimation of the CDF. First, in each stratum $h, (h = 1, \ldots, L)$, the PSU's are grouped at random

into two halves, h_1 and h_2, containing $m_{h_1} = [n_h/2]$ and $m_{h_2} = n_h - m_{h_1}$ PSU's, respectively. Setting the group indicator to

$$\delta_h^{(r)} = \begin{cases} 1, & h_1 \in r \\ -1, & h_2 \in r \end{cases}$$

where $r = 1, ..., R$ denotes a half-sample (replicate), the half-samples are balanced on the groups if $\sum_{r=1}^{R} \delta_h^{(r)} = 0$ and $\sum_{r=1}^{R} \delta_h^{(r)} \delta_{h'}^{(r)} = 0, (h \neq h')$. A minimal set of balanced half-samples can be obtained from a Hadamard matrix of order $R(L + 1 \leq R \leq L + 4)$.

The estimator of the distribution function based on the r-th half-sample is

$$\hat{F}^{(r)}(y) = \frac{\hat{G}^{(r)}(y)}{\hat{N}^{(r)}}$$

where

$$\hat{G}^{(r)}(y) = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci} I\{y_{hci} \leq y\}, \hat{N}^{(r)} = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}$$

and $A_{hc}^{(r)}$ is the weight modifier and is constant for all clusters in the same half-sample. We assume that the weights of all units (households) in a cluster are rescaled equally by the modifier $A_{hc}^{(r)}$.

The standard GBHS method, when n_h is even, uses

$$A_{hc}^{(r)} = \begin{cases} 1 + \delta_h^{(r)}, & c \in h_1, \\ 1 - \delta_h^{(r)}, & c \in h_2 \end{cases} \quad (3.3)$$

which means that the weights are modified either by 2 or 0 depending on whether a unit is in the replicate or not. When n_h is odd, a number of different modifications have been considered (see Shao 1993 and Sitter 1993).

The method that we are using is based on the standard balanced replication resampling plan and a variant of the rescaling method proposed by Shao (1993):

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - b_h)\delta_h^{(r)}, & c \in h_2. \end{cases}$$

The maintenance of the stratum sample size in any of the half-sample replicates means that

$$\sum_{c \in h1} [1 + (1 - a_h)\delta_h^{(r)}] + \sum_{c \in h2} [1 - (1 - b_h)\delta_h^{(r)}] = n_h,$$

which results in

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\,\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - a_h)\dfrac{m_{h_1}}{m_{h_2}}\,\delta_h^{(r)}, & c \in h_2. \end{cases} \quad (3.4)$$

To ensure the non-negativity of the modified weights, a_h should satisfy $0 \leq a_h < 1$. When n_h is even we would like (3.4) to reduce to (3.3). Following Shao's idea (1993), we want the GBHS variance estimator to agree with a consistent estimator of the variance in the case of linear statistics. This leads to the following requirements for the stratum-specific perturbation factors $1 - a_h$:

For all h: (i) $0 < 1 - a_h \leq 1$; (ii) $(1 - a_h)^2 (m_{h_1}/m_{h_2})^2 \approx 1$; (iii) $(1 - a_h)^2 m_{h_1}/m_{h_2} \approx 1$. For the even n_h's we simply let $1 - a_h = 1$. However, keeping $1 - a_h = 1$ for odd n_h's would exclude any contribution from the clusters in the first half-sample when $\delta_h^{(r)} = -1$, see equation (3.4). For the purpose of the simulation study we chose

$$1 - a_h = \sqrt{\frac{n_h}{2\,m_{h2}}} \quad (3.5)$$

which reduces to 1 for an even n_h. In the case of an odd stratum sample size it is equal to $\sqrt{1 - 1/(n_h + 1)}$. In our simulation study very few strata have an odd n_h and we obtain $v_{GB1}(\hat{\mu}_Y) = v_{GB2}(\hat{\mu}_Y) \approx v_L(\hat{\mu}_Y)$ where $\hat{\mu}_Y$ is the sample mean and $v_L(\hat{\mu}_Y)$ is the commonly used linearization variance estimator. However, it is felt that more research is needed into modifying the GBHS method to handle many strata containing an odd number of PSU's.

As in the case of the jackknife method, the estimate of the income inequality measure computed from the r-th half-sample is $\hat{\theta}^{(r)} = \sum_s J(\hat{F}^{(r)}, y_{hci}, \hat{\beta}^{(r)}) \tilde{w}_{hci}^{(r)}$ where $\hat{\beta}^{(r)}$ is an estimate of the nuisance parameter based on the r-th half-sample and $\tilde{w}_{hci}^{(r)} = \tilde{w}_{hci} A_{hc}^{(r)}$. The resulting GBHS variance estimator of $\hat{\theta}$ is

$$v_{GB1}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \quad (3.6)$$

By repeating the random grouping of units within each stratum T times, computing $v_{GB1}(\hat{\theta})$ each time and averaging over the T repetitions we obtain the Repeatedly Grouped Balanced Half Sample (RGBHS) variance estimator

$$v_{RG1}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} v_{GB1_t}(\hat{\theta}).$$

A variant of the GBHS estimator (and RGBHS) is obtained by replacing $\hat{\theta}$ by $\hat{\theta} = \sum_r \hat{\theta}^{(r)}/R$, and will be denoted by $v_{GB2}(\hat{\theta})$ (and $v_{RG2}(\hat{\theta})$).

Needless to say that when weights are calibrated they have to be properly modified for each GBHS replication using the same balanced half sample procedure.

3.3 Bootstrap Method

We also investigated the performance of the bootstrap method for variance estimation of different income statistics. We adopted the bootstrap resampling scheme for the stratified cluster sample as given by Rao, Wu and Yue (1992). Briefly, draw a simple random sample of $n_h - 1$ clusters with replacement (from the n_h sample clusters) independently in

Table 1
Definition of u^*_{hci} Variates for the EE Approach

Measure	u^*_{hci}
Gini Index	$2[\hat{A}(y_{hci})y_{hci} + \hat{B}(y_{hci}) - \hat{\mu}(\hat{G}+1)/2]/\hat{\mu}$ where $A(y) = \hat{F}(y) - \frac{\hat{G}+1}{2}$ and $B(y) = \sum_s w_{hcj} y_{hcj} I\{y_{hcj} \geq y\}$.
Lorenz Curve	$[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)]/\hat{\mu}$
Quantile Share	$\frac{1}{\hat{\mu}}[(y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\hat{Q}(p_1, p_2)]$
Quantile	$-[I\{y \leq \hat{\xi}_p\} - p]/\hat{f}(\hat{\xi}_p)$, $\hat{f}(.)$ is the finite population density estimator
Low Income Proportion	$-\frac{\hat{f}(\hat{\xi}_{0.5}/2)}{2\hat{f}(\hat{\xi}_{0.5})}[I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 1/2] + [I\{y_{hci} \leq \hat{\xi}_{0.5}/2\} - \hat{\Lambda}_{0.5}]$
Polarization Index	$\frac{2}{\hat{\xi}_{0.5}}[(\hat{\xi}_{0.5} - y_{hci})(I\{y_i \leq \hat{\xi}_{0.5}\} - 0.5) - (A(y_{hci})y_{hci} + B(y_{hci}) - (\hat{G}+1)\hat{\xi}_{0.5}/2 + \hat{G}y_{hci}/2)] + \frac{P\hat{I}}{\hat{\xi}_{0.5}\hat{f}(\hat{\xi}_{0.5})}(I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) - P\hat{I}$

each stratum. The bootstrap weight, w^*_{hci}, is obtained by modifying the original weight w_{hci} as follows:

$$w^*_{hci} = A_{hc} w_{hci}$$

where

$$A_{hc} = \frac{n_h}{n_h - 1} m^*_{hc}$$

and m^*_{hc} is the number of times the hc-th cluster is selected. Note that $\sum_c m^*_{hc} = n_h - 1$. This procedure is repeated independently B times; for each bootstrap sample, we calculate $\hat{\theta}^* = \sum_s J(\hat{F}^*, y_{hci}, \hat{\beta}^*)\tilde{w}^*_{hci}$ where $\hat{\beta}^*$ is an estimate of the nuisance parameter based on the bootstrap sample and $\tilde{w}^*_{hci} = w^*_{hci}/\sum_s w^*_{hci}$. The bootstrap estimate of the variance of $\hat{\theta}$ is then given by

$$v_{B1}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{(b)} - \hat{\theta})^2.$$

Another variance estimate is obtained by substituting $\hat{\theta}$ with the mean of bootstrap replicates.

3.4 Linearization via the Estimating Equations Approach

The estimating equations (EE) approach of Binder (Binder 1991, Binder and Patak 1994, Binder and Kovačević 1995), unlike the resampling methods, is not computationally intensive. This method, based on linearization, provides formulae for asymptotic variances which are easy to program despite their complicated appearance.

Applying the EE methodology as given in Binder and Patak (1994), Binder and Kovačević (1995) and Kovačević and Binder (1997) one obtains expressions for the approximate variance estimators of the studied measures as

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1}\sum_c \left(u^*_{hc} - \bar{u}^*_h\right)^2 \tag{3.7}$$

where $u^*_{hc} = \sum_i \tilde{w}_{hci} u^*_{hci}$, $\bar{u}^*_h = \sum_c u^*_{hc}/n_h$, and $\tilde{w}_{hci}$ is a normalized weight. For more on the EE approach, in particular the relationship between the u^*_{hci} variates and the J function, we refer the reader to Binder and Kovačević (1995). The u^*_{hci} variates for the considered measures are given in Table 1.

The expressions for the u^*_{hci} variates for the low income proportion and polarization index depend on the estimate of the density function at the median, $\hat{f}(\hat{\xi}_{0.5})$, and half of the median, $\hat{f}(\hat{\xi}_{0.5}/2)$. An appropriate method for estimating these quantities is given in Binder and Kovačević (1995).

4. SIMULATION STUDY

4.1 Data and the Design of the Simulation Study

The Ontario sample from the 1988 Canadian Survey of Consumer Finance (SCF) was used as the underlying population of the study. The SCF is an annual supplement to the monthly Canadian Labour Force Survey. The population contained 7474 households in 525 PSU's from 40 strata. Originally, the Ontario sample was taken from 91 strata which we collapsed to form sufficiently large strata. For each household a nonnegative value of the total annual income was available. The distribution of the income on this micro population was highly skewed to the right with coefficients of skewness and kurtosis obtained as 4.5 and 89.5, respectively. The true values of the parameters of interest (measures of income inequality and polarization) were computed from this population. Neyman allocation was used to assign 108 sample clusters (PSU's) to the 40 strata. A one-stage cluster design with the strata samples sizes between 2 and 6 clusters, selected with probability proportional to size and with replacement was used. In a selected cluster all households (6 to 20) were enumerated.

We considered the following measures in the study: Gini Index, Low Income Proportion, Polarization Index, a set of

Quantile Shares, a set of Lorenz Curve Ordinates and the corresponding quantiles. The MSE's of the estimates of these measures were approximated by the empirical mean squared error (EMSE), computed over 10,000 independent samples drawn by the design explained above. These EMSE's were used as 'true' MSE's for comparison with the estimated variances.

From each of the 10,000 samples, along with the estimates of the parameters, we computed estimates of the sampling variances using the following methods: the delete-one-PSU jackknife (JK), the grouped balanced half-sample (GBHS) and the repeatedly grouped balanced half-sample (RGBHS), the bootstrap (BS) and the linearization method via estimating equations (EE). For all resampling methods two different estimators were used, one using the 'full sample' estimate and another one using the mean over all replicates. The jackknife variance estimators were based on 108 jackknife replicates while the bootstrap method was based on 100 replicates. The GBHS and RGBHS were based on 44 balanced replicates obtained from a 44 by 44 Hadamard matrix and 3 repetitions for RGBHS, totalling 132 half-sample replicates for this method. Note that the number of jackknife replicates is non-arbitrary and is determined by the number of clusters in the sample. Similarly, the number of GBHS replicates is determined by the number of strata. In order to make the number of replicates comparable over all methods, we decided to have 100 ($\approx$ 108) bootstrap replicates and 3 repetitions of the GBHS resulting in 132 replicates for RGBHS.

In order to evaluate the accuracy and the precision of the considered methods we computed their relative biases and relative variance (instability) over the $A = 10{,}000$ simulations:

$$\text{rel. bias}(v_M) = \frac{\sum_a v_M(a)/A - \text{EMSE}}{\text{EMSE}}$$

$$\text{rel. var.}(v_M) = \frac{\sqrt{\sum_a [v_M(a) - \text{EMSE}]^2/A}}{\text{EMSE}}$$

To evaluate the effectiveness of normal-theory confidence intervals, empirical coverage rates were computed for nominal confidence coefficients of $100(1-\alpha)\% = 90, 95$ and 99 percent,

$$\text{cov. prob.}(v_M) = \frac{\sum_a I\{|\hat{\theta}_a - \theta| / \sqrt{v_M(a)} \le z_{\alpha/2}\}}{A}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-th standard normal percental. Upper and lower tailed error rates were also calculated as follows,

$$\text{err\_}L(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) / \sqrt{v_M(a)} < -z_{\alpha/2}\}}{A}$$

$$\text{err\_}U(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) / \sqrt{v_M(a)} > z_{\alpha/2}\}}{A}.$$

The large set of results obtained from the simulation study are summarized separately for each income inequality measure.

4.2 Summary of Findings

Gini Index

Concerning the accuracy of the variance estimators for the Gini index, all methods performed similarly, with very small negative relative biases ranging between -2.2 and -0.6 percent. Of all the estimators, the RGBHS estimators had the smallest relative bias.

All estimators were found to be of approximately the same stability, in the range of 87-99%. The grouped balanced half-sample methods (GBHS and RGBHS) perform slightly worse than other methods.

The coverage probabilities for the 95% confidence intervals were in the range of 92.6 (for GBHS) to 93.9 (for RGBHS). The lower tail error rates were understated by the nominal 2.5% rate for all methods considered. We found that the lower tails were more than 100% heavier than the nominal 2.5%, ranging between 4.6 and 5.4%. The upper tail error rates were overstated by the nominal rate for all methods. (See Table 2). We also computed the coverage rates for the 90% and 99% confidence intervals and they were in the range of 87.2 (for GBHS) to 88.5 (for RGBHS) and in the range of 97.7 (for GBHS) to 98.5 (for RGBHS), respectively. Similarly, the tail rates for the nominal 5% and 1% followed the pattern of 2.5%.

Overall, for variance estimation of the Gini index it is difficult to say which method is the best since all compared methods performed similarly. There is a slight trade off between accuracy and stability in the case of the balanced half-sample methods which give the most accurate estimates of the variance but at the same time the least stable. The empirical coverage probabilities for all of the estimators are also very similar. The realized values of the tail error rates suggest that the use of asymmetric confidence intervals is more appropriate.

Low Income Proportion (LIP)

All methods considered tended to overestimate the variance of the LIP. However, the difference in the magnitude of overestimation was large, and ranged between 1.1% for the EE and 76.9% for the JK1. The best performer among resampling methods was the bootstrap, where the relative bias for the BS1 estimator was 8.9% and for BS2 3.8%.

The jackknife estimate of the variance of the LIP was very unstable. The GBHS estimators also had increased instability. The bootstrap and EE estimators performed similarly with relative variation between 31 and 45%.

Table 2
Values of the Evaluation Statistics for the Variance Estimators of the Gini Index

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		-1.3	-1.3	-0.9	-1.1	-0.6	-0.7	-1.2	-2.2	-1.5
Relative Variation (%)		87.1	87.1	99.4	99.2	95.2	95.1	88.5	87.6	87.0
Coverage Probability (95%)		93.8	93.8	92.6	92.6	93.9	93.9	93.5	93.4	93.7
Tail Error Rates (2.5%)	L	4.8	4.8	5.4	5.4	4.6	4.6	5.0	5.1	4.9
	U	1.4	1.4	2.0	2.0	1.5	1.5	1.5	1.5	1.4

Table 3
Values of the Evaluation Statistics for the Variance Estimators of the Low Income Proportion

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		76.9	58.4	25.8	21.0	26.8	21.9	8.9	3.8	1.1
Relative Stability (%)		113.1	81.0	62.5	61.0	40.8	39.5	35.1	33.5	31.0
Coverage Probability (95%)		97.4	96.9	94.6	94.1	96.2	95.7	93.9	93.3	93.2
Tail Error Rates (2.5%)	L	2.1	2.6	3.3	3.5	2.4	2.6	4.6	5.0	5.0
	U	0.5	0.6	2.0	2.4	1.4	1.7	1.5	1.7	1.7

The 95% confidence interval for the LIP based on the JK variance estimates had higher than nominal coverage rates, 97.4 and 96.9%, consequences of the overestimation of the variance. The other methods had slightly lower coverage rates than nominal. The tail error rates showed that all methods resulted in heavier lower tails, indicating a skewed distribution of the LIP with a long tail to the right. For the cases of 90% and 99% confidence intervals we obtained exactly the same pattern for the coverage and the tail error rates.

Overall, for variance estimation of the LIP, the bootstrap and the EE method show supremacy over the other methods considered.

Polarization Index

The evaluation statistics for the variance estimators of the polarization index showed a high level of agreement in performance with variance estimation for the low income proportion. Again, the bootstrap and EE method were the best.

Table 4
Values of the Evaluation Statistics for the Variance Estimators of the Polarization Index

		Jackknife		GBHS		RGBHS		Bootstrap		Estimating Equations
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Relative Bias (%)		95.4	56.5	13.9	11.2	14.7	12.1	6.0	2.9	4.2
Relative Stability (%)		138.7	78.5	77.5	75.9	60.0	58.6	48.4	47.0	50.0
Coverage Probability (95%)		98.6	98.0	94.2	93.8	95.4	95.2	95.0	94.7	94.4
Tail Error Rates (2.5%)	L	0.7	0.8	2.2	2.4	1.4	1.4	1.8	2.0	2.0
	U	0.8	1.1	3.6	3.9	3.2	3.4	3.2	3.4	3.6

Lorenz Curve Ordinates and Quantile Shares

The full results for the Lorenz Curve Ordinates and Quantile Shares are given in Kovačević, Yung and Pandher (1995). We present here a graphical summary of the results in Figures 1a-1c. The jackknife method (both estimators) significantly overestimates the variances of all considered Lorenz Curve Ordinates (LCO) and Quantile Shares (QS). The relative bias of the JK1 estimator for the LCO ranged between 15 and 45% and between 9 and 27% for the JK2 estimator. The relative bias was smaller in the middle of the interval ($0 \le p \le 1$) and almost three times larger at the tails (for small and large values of p). The relative bias of the JK1 estimator was about 50% larger than the relative bias of the JK2 estimator for the LCO. The difference can be attributed to the significant difference between the full sample estimate of the LCO and the average taken over jackknife replicates.

Similar findings held for the performance of the JK variance estimators for QS's which overestimated the variance between 26-237%, depending on the population share. The largest overestimation appeared in the middle. Again, the JK1 was larger than JK2 by about 75%.

The magnitude of the relative bias was very small for the other two methods. However, there was no clear pattern about the direction of bias – sometimes it was positive, but often it was negative. The bootstrap estimators and the EE estimator outperformed the other methods, especially around the LCO corresponding to $p = 0.5$ (see Figure 2a). For clarity of the graphical presentation the JK methods are not shown in Figures 2a and 2b.

The variance of the QS's is estimated similarly. The bootstrap and EE provided the most accurate estimates of the variances of LCO and QS. For the LCO the relative bias ranged between -2 and $+3\%$ for bootstrap and -5 to $+1\%$ for EE. At the same time, for the QS, the bootstrap estimates had relative biases between -3 and $+8\%$ and EE estimates between -3 and $+5\%$.

Concerning the stability of the different variance estimators we found that all methods perform similarly with a slight advantage for the EE method. Also, there is an obvious direct dependence of the relative variation measure and the value of p.

When we compared the methods according to the coverage properties of the variance estimators for the LCO and QS we found that for the nominal 95% confidence interval, the JK method gave empirical coverage rates between 94.5 and 96.5% for the LCO and 94.5 to 99% for the QS. Other methods performed similarly with coverage rates between 88 and 94%. Better coverage was found for the LCO and QS with smaller value of p (see Figure1c). In contrast to findings for the Gini index, the lower tail error rates were about twice the upper tail error rates for all methods and for both LCO and QS. A similar pattern was observed for 90% and 99% confidence intervals.

Our empirical findings suggest that the jackknife method is not a good choice for the variance estimation of the LCO and QS especially for small and large values of p. Much better alternatives are the GBHS or the RGBHS. However, the best choice is either the EE method or the bootstrap.

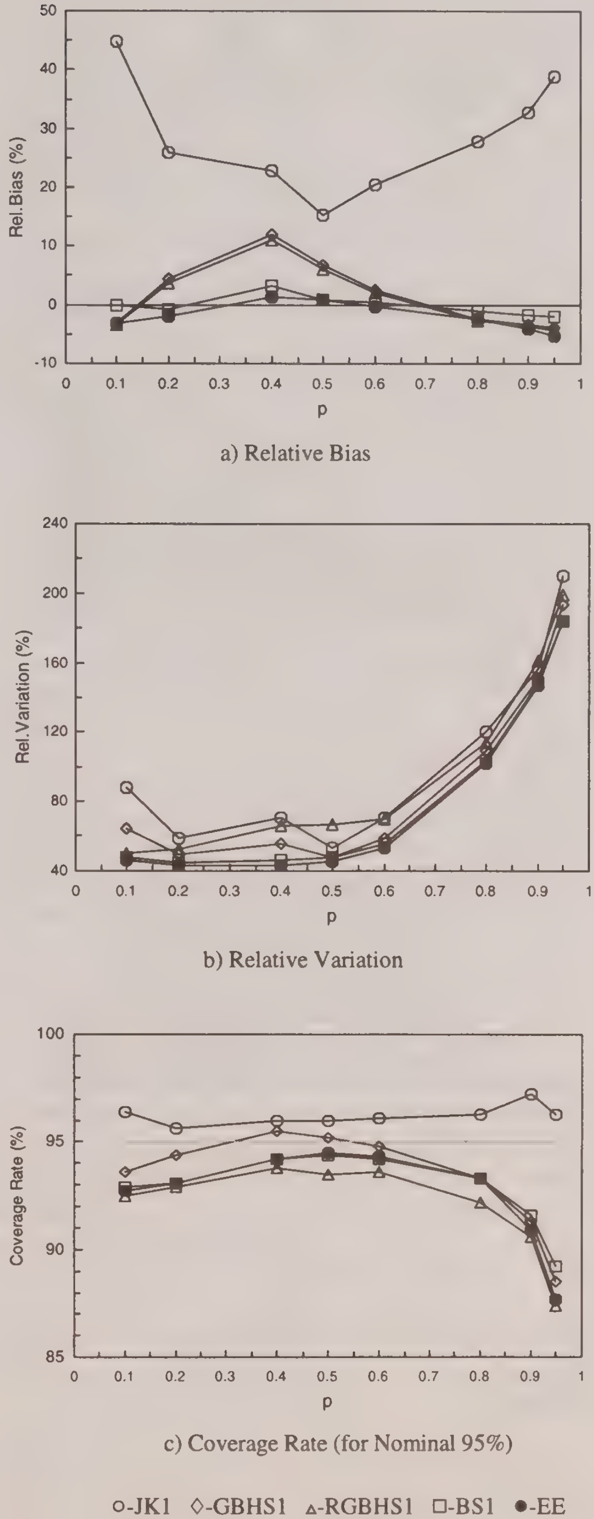
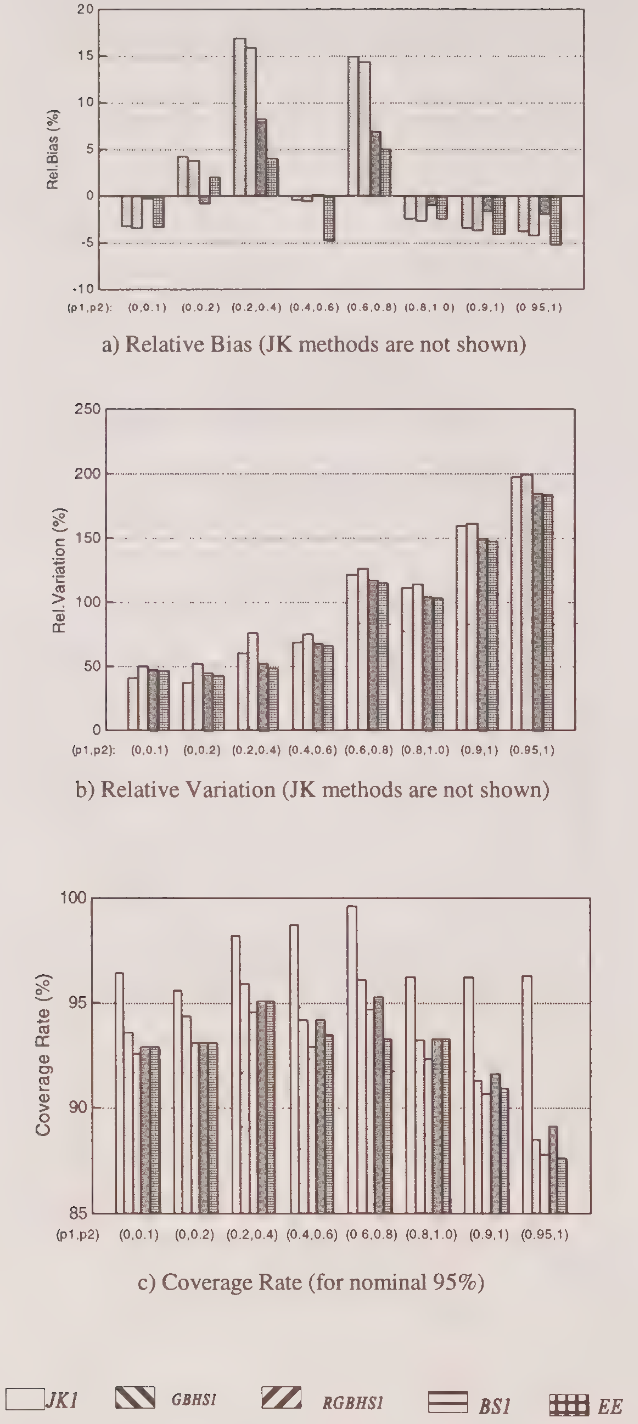


Figure 1. Properties of the Variance Estimators of Lorenz Curve Ordinates



a) Relative Bias (JK methods are not shown)

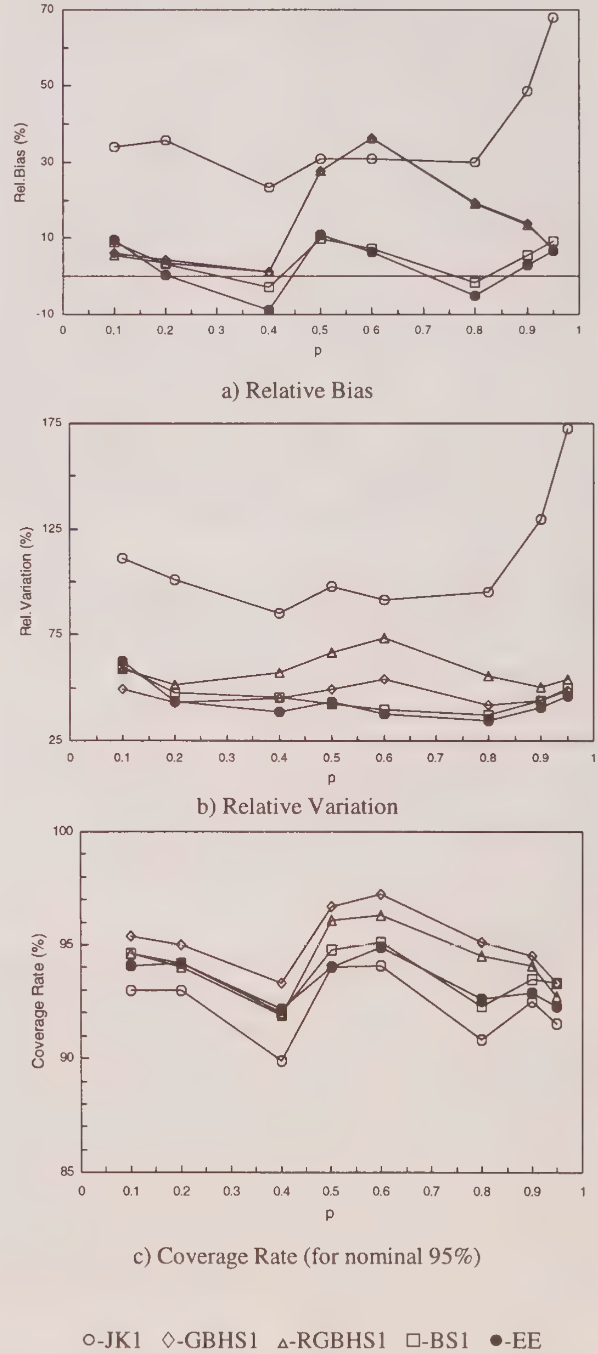
b) Relative Variation (JK methods are not shown)

c) Coverage Rate (for nominal 95%)

Figure 2. Properties of the Variance Estimators of Quantile Shares

Quantiles

The full results obtained for the quantiles are presented in Kovačević, Yung and Pandher (1995) and are summarized graphically here. The relative bias of the JK1 estimate of the variance for the quantiles was between 23 and 67% and for JK2 between 17 and 52%. The largest overestimation occurred for the variances of $\hat{\xi}_{0.90}$ and $\hat{\xi}_{0.95}$. The RGBHS and GBHS show quite a different picture. The variance of the median was overestimated by 27% but the variances of tail quantiles were obtained very accurately, with the relative bias between 3 and 7%. Other methods also performed much better for the tail quantiles and moderately better for the median and quantiles around it. In particular, the bootstrap and the EE method produced estimates with the smallest relative biases, although without clear pattern about the direction of the bias. For the bootstrap estimators, the relative bias was in the interval (−5%,+9%), and for EE (−8%, +9%) (see Figure 3a).



a) Relative Bias

b) Relative Variation

c) Coverage Rate (for nominal 95%)

Figure 3. Properties of the Variance Estimators of Quantiles

Table 5
Rankings of methods by relative bias, relative stability and empirical coverage probability

	Jackknife	GBHS	RGBHS	Bootstrap	EE (Taylor)	Best methods
Gini Index	All procedures performed similarly					–
Quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	EE, BS
Lorenz Curve	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	EE, BS
Quantile Shares	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 2	2, 1, 3	BS, EE
Low Income	5, 5, 5	3, 4, 2	4, 3, 1	2, 2 ,3	1, 1, 4	EE, BS
Polarization Index	5, 5, 5	3, 4, 4	4, 3, 2	2, 1, 1	1, 2, 3	BS, EE

The jackknife estimators were the least stable. The RGBHS, bootstrap and EE showed similar stability which, on average over all quantiles, was about three times higher than the stability of JK estimators. The highest stability was attained around the median (see Figure 3b).

In general, the coverage probabilities for the quantiles were less than nominal for all of the methods considered, with some exceptions for the GBHS and RGBHS methods (see Figure 3c). When we compared the observed tail error rates, it seemed that all methods exhibited similar behaviour, for the lower quantiles ($p = 0.1, 0.2$) the upper (right) tails were heavier; for others it was opposite, the lower tails were heavier. Similar results were obtained for the 90% and 99% confidence intervals.

The findings from this empirical study confirm that for variance estimation of quantiles, the jackknife method should be avoided. For the variance of the median, in particular, the best choice seems to be either the EE or the bootstrap. For other quantiles the RGBHS showed very good performance as well.

We condense our findings in Table 5 where the relative bias, relative variation and the coverage probabilities for the methods considered were ranked from 1 to 5 (1 = the best). For the resampling methods we averaged the values over both estimators. For the quantiles, LCO and QS we averaged the values over all p's. The last column contains the choice of the two best performing methods.

5. DISCUSSION AND CONCLUSION

The linearization method via EE has shown the best overall performance, the smallest relative bias, the smallest relative variation and relatively good coverage properties. Next to the EE method is the bootstrap method, as the best resampling method considered. The RGBHS and GBHS method performed comparably well for the Lorenz Curve ordinates, quantile shares and some of the quantiles, in the sense of the small relative bias and relative stability comparable with the bootstrap method. The jackknife method has performed poorly for all measures except the Gini index.

It is well known that the jackknife variance estimator performs poorly for non-smooth functions. The smoothness of the J function defined in (3.1) is an essential determinant of the asymptotic properties of its variance estimator. Classifying our measures as smooth or non-smooth on the basis of the J functions, we see that the only smooth estimator considered here was the Gini index. Not surprisingly, the Gini index was the only measure for which the jackknife performed well. However, when considering the jackknife variance estimator, care must be taken to ensure that the assumptions under which the jackknife is valid are fulfilled.

If the goal is to provide one method for variance estimation for the large list of different income statistics, our empirical study has shown that the bootstrap is the best resampling choice, and that the linearization via the estimating equations approach is the best computationally non-intensive method, which however, requires some preparatory algebraic work, different for each measure.

It should be emphasized that the empirical study was based on an one-stage cluster sampling design, with the clusters selected proportionally to their size, so the intracluster variability was not accounted for. Some other limited studies have shown similar behaviour of these methods in the case of two stage sampling plans (see Binder and Kovačević 1995, and Kovačević and Binder 1997).

ACKNOWLEDGMENTS

The authors would like to thank G.S. Pandher for his fruitful participation at the beginning of the project, J. Gambino for his thorough reading of an earlier version of the paper, H. Mantel, associate editor, anonymous referees and the editor, for valuable comments that significantly improved quality of the paper.

REFERENCES

BABU, G.J. (1986). A note on bootstrapping the variance of sample quantile. *Annals of the Institute of Statistical Mathematics*, 38-A, 439-443.

BEACH, C.M., and DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BEACH, C.M., and KALISKI, S.F. (1986). Lorenz curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics*, 35, 38-45.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., and KOVAČEVIĆ, M.S. (1995). Estimating some measures of income inequality from survey data: an application of the estimating equation approach. *Survey Methodology*, 21, 137-145.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

EFRON, B. (1979). Bootstrap method: another look at the jackknife. *Annals of Statistics* 7, 1-26.

FOSTER, J.E., and WOLFSON, M.C. (1992). Polarization and the decline of the middle class: Canada and the U.S. (Manuscript).

FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KAKWANI, N.C. (1980). *Income Inequality and Poverty*. Washington, D.C.: World Bank.

KOVAČEVIĆ, M.S., and BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization-the estimating equations approach. (To appear in *Journal of Official Statistics*).

KOVAČEVIĆ, M.S., YUNG, W., and PANDHER, G.S. (1995). Estimating the Sampling Variances of Income Inequality and Polarization – An Empirical Study. Methodology Branch Working Paper, HSMD-95-007-E. Statistics Canada.

KOVAR, J.G. (1987). Variance Estimation of Medians in Stratified Samples. Methodology branch working paper, BSMD-87-004-E. Statistics Canada.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics* 16, 25-45.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LOVE, R., and WOLFSON, M.C. (1976). Income inequality: statistical methodology and Canadian illustrations. Ottawa, Statistics Canada.

McCARTHY, P.J. (1993). Standard error and confidence interval estimation for the median. *Journal of Official Statistics*, 9, 673-689.

NYGÄRD, F., and SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist & Wiksell International.

NYGÄRD, F., and SANDSTRÖM, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 399-412.

RAO, J.N.K., and SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from survey data: Some recent work. *Proceedings of the 46th Session of International Statistical Institute*, 3, 5-19.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

SANDSTRÖM, A., WRETMAN, J.H., and WALDÉN, B. (1985). Variance estimators of the Gini coefficient, simple random sampling. *Metron*, 43, 41-70.

SANDSTRÖM, A., WRETMAN, J.H., and WALDÉN, B. (1988). Variance estimators of the Gini coefficient – probability sampling. *Journal of Business and Economic Statistics*, 6, 113-119.

SEN, A.K. (1973). *On Economic Inequality*. London: Oxford University Press.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 119-122.

SHAO, J. (1993). Balanced repeated replication. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 544-549.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

WOLFSON, M.C. (1994). When inequalities diverge. *American Economic Review*, 84, 353-358.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

YITZHATI, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics*, 9, 235-239.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 53-60
Statistics Canada

Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error

K. HUMPHREYS and C. J. SKINNER[1]

ABSTRACT

The problem of estimating transition rates from longitudinal survey data in the presence of misclassification error is considered. Approaches which use external information on misclassification rates are reviewed, together with alternative models for measurement error. We define categorical instrumental variables and propose methods for the identification and estimation of models including such variables by viewing the model as a restricted latent class model. The numerical properties of the implied instrumental variable estimators of flow rates are studied using data from the Panel Study of Income Dynamics.

KEY WORDS: Latent class; Longitudinal; Misclassification; Transition rate.

1. INTRODUCTION

One of the major benefits of longitudinal surveys is that they permit the estimation of gross flows, for example flows out of unemployment into employment (see *e.g.*, Hogue and Flaim 1986). A key problem when estimating flows is the bias induced by measurement error. For the estimation of cross-sectional proportions, misclassification into and out of states may tend to cancel out (Chua and Fuller 1987). Such compensation tends not to occur, however, when estimating longitudinal flows.

The first response to the problem of measurement error should clearly be to attempt to reduce the error in the survey measurement procedures. Relevant approaches are discussed by Biemer, Groves, Lyberg, Mathiowetz and Sudman (1991), but will not be considered here. Even with the "best" survey procedures, however, some measurement error will inevitably arise and there will remain a need to compensate for the effect of error in the survey analysis.

Methods for compensating for measurement error are generally based on some assumed model of the error process. Some models which have been proposed in the literature will be referred to in Section 2. In order to identify and estimate these models it is generally necessary to use additional auxiliary information, such as provided by reinterview studies (*e.g.*, Meyer 1988). Since reinterview studies are costly, however, and since in practice their aim is often not to estimate the characteristics of the measurement error distribution (Forsman and Schreiner 1991), there remains a need for alternative procedures which may be used when no reinterview data is available. For measurement error on continuous variables, a common approach employed in the absence of auxiliary information about the measurement error distribution is the method of instrumental variable estimation (*e.g.*, Fuller 1987, Sect. 1.4). An instrumental variable is a variable included in the survey dataset which is related to the true variable measured with error but is uncorrelated with the measurement error. These and associated assumptions supply information which replaces that provided by reinterview studies and enables parameters of the model involving the true variable to be identified and estimated. The aim of this paper is to investigate how the instrumental variable estimation method may be adapted to estimate flows among discrete states. We find that latent class models (*e.g.*, Bartholomew 1987, Ch. 2) provide a general framework within which the assumptions about the instrumental variable correspond to certain restrictions on the model parameters. Our approach is thus related to other approaches which impose restrictions on latent class models (*e.g.*, van de Pol and de Leeuw 1986; van de Pol and Langeheine 1990).

2. MODELS

We consider only the case of two occasions $t = 1$ and $t = 2$. Let the number of states into which each individual can be classified at each occasion be r. Denote the classified states at $t = 1$ and $t = 2$ by X and Y respectively and the corresponding true states by x and y. We assume a model in which the vectors of values of (X, Y, x, y) are generated as independent outcomes of a common random vector with distribution $\mathrm{pr}(X = i, Y = j, x = u, y = v)$.

The first assumption about this distribution, made by a number of authors (*e.g.*, Abowd and Zellner 1985; Poterba and Summers 1986 and Chua and Fuller 1987) and which we shall also make, is that the classification errors on the two occasions are conditionally independent given the true states, that is

$$\mathrm{pr}(X = i, Y = j \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u, y = v)\, \mathrm{pr}(Y = i \mid x = u, y = v). \quad \text{(A1)}$$

[1] K. Humphreys, Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden; C.J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

Such an assumption is common in general latent variable models (*e.g.*, Anderson 1959). It seems a reasonable initial assumption when the survey measurement procedures are independent on the two occasions. On the other hand, if X is obtained retrospectively from the same interview in which Y is measured then it seems likely that the tendency for respondents to give over-consistent responses in a single interview may tend to induce positive association between classification errors. See, for example, Marquis and Moore (1990) on evidence from the Survey of Income and Program Participation. A further reason for doubting the conditional independence assumption is the possibility of individual heterogeneity in misclassification probabilities, for example some respondents may be more reliable than others. See Skinner and Torelli (1993) and Singh and Rao (1995). In Section 4 we shall allow for heterogeneity by assuming only that the model holds within cells of a cross-classification of observed variables.

Our next basic assumption is that classification error only depends on current true state so that

$$\mathrm{pr}(X = i \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u) = K_{xiu}, \text{ say,}$$

$$\mathrm{pr}(Y = j \mid x = u, y = v) = \mathrm{pr}(Y = j \mid y = v) = K_{yjv}, \text{ say.} \quad \text{(A2)}$$

The K_{xiu} and K_{yjv} define $r \times r$ misclassification matrices $\boldsymbol{K}_x = [K_{xiu}]$ and $\boldsymbol{K}_y = [K_{yjv}]$. Letting $\boldsymbol{P}$ denote the $r \times r$ matrix with ij-th element $\mathrm{pr}(X = i, Y = j)$ and Π the $r \times r$ matrix with uv-th element $\mathrm{pr}(x = u, y = v)$ we have the matrix equation

$$\boldsymbol{P} = \boldsymbol{K}_x \Pi \boldsymbol{K}_y'. \quad (1)$$

The matrix Π contains the parameters of interest, whereas it is the matrix $\boldsymbol{P}$ which may be estimated consistently from sample X and Y values. If auxiliary estimates of $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ are available and these are non-singular then we can solve equation (1) to obtain estimates of Π. If it is possible to ascertain the true states in reinterview studies then $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ may be estimated directly (Abowd and Zellner 1985). On the other hand, if the reinterview study only provides independent reclassifications then it is only possible to estimate the interview-reinterview matrices

$$\boldsymbol{K}_x \Delta_x \boldsymbol{K}_x' \text{ and } \boldsymbol{K}_y \Delta_y \boldsymbol{K}_y'$$

where $\Delta_x = \mathrm{diag}[\mathrm{pr}(x = u)], \Delta_y = \mathrm{diag}[\mathrm{pr}(y = v)]$ (Chua and Fuller 1987). Each interview-reinterview matrix is symmetric with elements summing to one and so only contains $r(r+1)/2 - 1$ "independent" items of information. Since each column of each $\boldsymbol{K}$ matrix and the diagonal of each Δ matrix sum to one, the number of unknown parameters on each occasion is $r(r-1) + r - 1 = r^2 - 1$. The excess of parameters over items of information is therefore $r^2 - 1 - r(r+1)/2 + 1 = r(r-1)/2$ at each occasion and so the model is underidentified for $r \geq 2$. Chua and Fuller (1987) suggest that a natural extra assumption to make to help achieve identification is to suppose that the measurement errors are unbiased on each occasion in the sense that

$$\mathrm{pr}(x = i) = \mathrm{pr}(X = i), \ \mathrm{pr}(y = i) = \mathrm{pr}(Y = i) \ \ i = 1, \ldots, r. \quad (2)$$

In this case false positives and false negatives tend to compensate for each other in cross-sectional estimates of proportions. This assumption reduces the number of parameters by $r - 1$ on each occasion. Even under this assumption the model remains underidentified for $r \geq 3$ and Chua and Fuller (1987) have to introduce further assumptions.

Let us now consider how the model might be identified when no reinterview data is available. For simple linear regression with measurement error in the covariate, the instrumental variable approach (Fuller 1987, Sect. 1.4) assumes the availability of an observed "instrumental" variable W, which is correlated with the covariate, but is independent of the measurement error and independent of the error in the regression equation. We extend this assumption to our framework by defining W to be an *instrumental variable* if it is not independent of x and if

W and (X, Y) are conditionally independent given (x,y), (A3)

W and y are conditionally independent given x. (A4)

In general we shall allow W to be a categorical variable with an arbitrary number s of categories, although since we shall desire W to be closely related to x, we shall usually have $s = r$ in practice. One specific possibility is to take W as the classified state at time $t - 1$. This use of a lagged value of a "covariate" as an instrumental variable may be traced back to the earliest discussions of instrumental variable estimation (*e.g.*, Reiersol 1941; Durbin 1954). In this case, assumption A4 follows if the true states obey a Markov process and the classification errors are conditionally independent, as in A1.

The model resulting from assumptions (A1)-(A4) may be represented by the conditional independence graph in Figure 1. Each vertex in the graph represents a variable. Edges between pairs of vertices are absent if the corresponding variables are conditionally independent given the remaining variables.

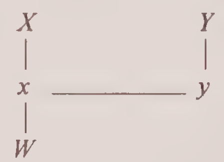


Figure 1. Conditional Independence Graph of Basic Model

The model is an example of a restricted latent class model (Goodman 1974), where the observed variables X, Y and W are conditionally independent given the latent variables x and y, that is they are independent within the r^2 latent classes defined by the pairs of values of (x,y). There are $2(r-1)r^2 + (s-1)r^2 + (r^2 - 1)$ parameters of this model given by the $(r-1)r^2$ parameters $\mathrm{pr}(X = i \mid x = u, y = v)$, the $(r-1)r^2$ parameters $\mathrm{pr}(Y = j \mid x = u, y = v)$, the $(s-1)r^2$ parameters $\mathrm{pr}(W = k \mid x = u, y = v)$ and the $r^2 - 1$ free

parameters $\mathrm{pr}(x=u, y=v)$. These parameters are subject to the $2r(r-1)^2$ restrictions in (A2) and the $(s-1)r(r-1)$ restrictions implied by (A4). We first restrict attention to the case $r=2$. In this case there are $4s+7$ parameters subject to $2s+2$ restrictions, leaving $2s+5$ free parameters

$$\{K_{x2u}, K_{y2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u=1,2\ v=1,2\},$$

where $\varphi_{ku} = \mathrm{pr}(W=k \mid x=u)$, $\theta_u = \mathrm{pr}(y=2 \mid x=u)$, and $\pi = \mathrm{pr}(x=2)$. The number of "free" cell probabilities in the observed table of X by Y by W is $r^2 s-1$, or $4s-1$ when $r=2$. Hence a necessary condition for identification when $r=2$ is that $4s-1 \geq 2s+5$ or $s \geq 3$. Unfortunately, this is not a sufficient condition. For let

$$R_u = \mathrm{pr}(Y=2 \mid x=u) = \sum_{v=1}^{2} K_{y2v}\,\theta_u^{v-1}(1-\theta_u)^{2-v}. \tag{3}$$

Then

$$\mathrm{pr}(X=i, Y=j, W=k) = \sum_{u=1}^{2} K_{xiu}\varphi_{ku} R_u^{j-1}(1-R_u)^{2-j}\pi^{u-1}(1-\pi)^{2-u}. \tag{4}$$

Hence the $4s-1$ free cell probabilities are determined by just the $2s+3$ parameters

$$\{K_{x2u}, \varphi_{2u}, \ldots, \varphi_{su}, R_u, \pi;\ u=1,2\}$$

so a necessary condition for identification of these parameters is that $4s-1 \geq 2s+3$ or $s \geq 2$. In fact this is also a sufficient condition for identification of these parameters, except for certain exceptional combinations of these parameters. (See Madansky (1960) for the case $s=2$ and Goodman (1974) for the case of general $s \geq 2$.)

However, even though the above $2s+3$ parameters are in general identified for $s \geq 2$ it is not possible to determine the 4 parameters $K_{y21}, K_{y22}, \theta_1$ and θ_2 since they are related to only two identified parameters, R_1 and R_2, via equation (3). In particular the key parameters of interest θ_1 and θ_2 remain underidentified whatever the value of s.

It is therefore necessary to impose at least 2 further restrictions on the model to identify θ_1 and θ_2. Following Chua and Fuller (1987), one idea would be to assume unbiased measurement errors as in (2) which imposes the two constraints

$$\pi = K_{x21}(1-\pi) + K_{x22}\pi \tag{5}$$

$$\theta_1(1-\pi) + \theta_2\pi = R_1(1-\pi) + R_2\pi. \tag{6}$$

Unfortunately the first constraint only applies to the parameters which are already identified for $s \geq 2$ so these constraints are insufficient to identify θ_1 and θ_2. An alternative assumption which we shall make is that the error process is constant over time so that

$$K_{xiu} = K_{yiu} = K_{iu}, \quad \text{say,} \quad \text{for } i,u = 1,2,\ldots,r. \tag{A5}$$

This seems a natural basic assumption if the same survey measurement procedure is used over time. The under-identification problem for the case $r=2$ discussed above is removed by this assumption since, given the identification of $K_{xiu} = K_{iu}$ and R_u, we can determine θ_u from (3) by

$$\theta_u = (R_u - K_{21})/(K_{22} - K_{21}) \tag{7}$$

(excluding the trivial case when the measured variables are independent of the true variables so that $K_{22} = K_{21}$).

In summary, when assumptions (A1) - (A5) hold and $r=2$, our model has $2s+3$ free parameters $\{K_{2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u=1,2\}$ which are identified if $s \geq 2$, except in exceptional cases such as discussed by Madansky (1960).

Finally, let us return to the case of general r. Since (A5) imposes $(r-1)r$ restrictions, the number of free parameters becomes $2(r-1)r^2 + (s-1)r^2 + (r^2-1) - [2r(r-1)^2 + (s-1)r(r-1)] - (r-1)r = 2r^2 + sr - 2r - 1$. There are $r^2 s - 1$ free cell probabilities in the table of X by Y by W so the model will in general be identified if $r(r-1)(s-2) \geq 0$. Thus the condition for identification of these parameters remains $s \geq 2$, for any value of $r \geq 2$. Furthermore we can write

$$R_{ju} = \Pr(Y=j \mid x=u) = \sum_{v=1}^{r} K_{jv}\theta_{uv}$$

where $\theta_{uv} = \mathrm{pr}(y=v \mid x=u)$. Hence, provided the matrix $[K_{iu}]$ is non-singular, the θ_{uv} may be determined from the R_{ju} and K_{jv} and hence are also identified. Thus for general r, the model is identified under assumptions (A1)-(A5), except for exceptional cases as discussed by Goodman (1974).

3. ESTIMATION

We shall suppose that for a sample of size n we observe counts n_{ijk} in the cells of the $r \times r \times s$ contingency table of $X \times Y \times W$, and that these are multinomially distributed with parameters n and $p_{ijk} = \mathrm{pr}(X=i, Y=j, W=k)$. The implied log likelihood is

$$l = \sum_i \sum_j \sum_k n_{ijk} \log p_{ijk}.$$

Under a complex sampling design, we may take the n_{ijk} to be weighted counts, giving a pseudo log likelihood (Skinner 1989). The estimators of the parameters obtained by maximising l will be called *instrumental variable* (IV) estimators.

For the remainder of this paper we shall only consider the case $r=s=2$ when the model is just identified (except for exceptional values of the parameters). In this case we might

attempt to set $p_{ijk} = n_{ijk}/n$ and then solve equations (6) and (7) for the unknown parameters. If the resulting solutions lie within the feasible parameter space, that is probabilities lie in the range [0,1], then these solutions will be the IV estimates. However, in practice we have found that, for moderate sample sizes, infeasible solutions can often arise. Furthermore the solution of these equations is not computationally straightforward. Hence we have found it easier to maximise l directly using the numerical procedures in the package GAUSS (Edlefsen and Jones 1984) or else by using packages which fit latent class models using the EM algorithm such as PANMARK (van de Pol, Langeheine and de Jong 1991). For a latent class package it would be possible to fit an unrestricted two class model and then to estimate θ_1 and θ_2 via (7). However, there would be no guarantee that the resulting estimates would lie in the feasible range [0,1] with this approach. Furthermore there would be the additional complication of determining standard errors for the estimates of θ_1 and θ_2 from the covariance matrix of the estimates of $(R_1, R_2, K_{21}, K_{22})$. Hence we have found it more convenient to fit the model directly as a restricted latent class model. A further advantage of this approach is that it extends naturally to the fitting of similar models across subgroups subject to possible constraints that some parameters are constant across subgroups. This possibility is explored further in Section 4.

Under multinomial assumptions, standard errors may be based on the second derivatives of the log-likelihood evaluated at the IV estimates. This approach becomes problematic, however, if the maximum of l is at the boundary of the parameter space. One approach then is simply to treat the values of the parameters at the boundary as known. However, this is likely to lead to underestimation of uncertainty. Baker and Laird (1988) consider two alternative approaches to obtaining interval estimates for individual parameters in such circumstances: a bootstrap method and a profile likelihood method. The bootstrap method involves drawing repeated multinomial samples with p_{ijk} set equal to n_{ijk}/n and recording the distribution of parameter estimates across repeated bootstrap samples. Interval estimates for given parameters are obtained by the profile likelihood methods as the sets of values of the parameter which are not rejected by a likelihood ratio test. These methods are illustrated at the end of Section 4.

4. NUMERICAL ILLUSTRATIONS

For the purpose of numerical illustration we use data from the equal probability subsample of the US Panel Study of Income Dynamics (PSID). See Hill (1992). We consider the two states employed and not employed, coded 1 and 2 respectively, thus restricting attention again to the binary variable case. For simplicity, we ignore non-response and consider the sample of 5,357 individuals aged 18-64 in 1986 with complete values on the variables: employment status in 1985, 1986 and 1987, car ownership, age, sex and education.

We assess the properties of the IV estimator in two ways. First, in Section 4.1, we compare the bias and standard error of the IV estimator with the "unadjusted" estimator for hypothetical instrumental variables, with a range of different associations with x. Second, in Section 4.2, we consider the impact of using different actual PSID variables as instrumental variables.

4.1 Bias and Standard Error Properties of Estimators for Hypothetical Instrumental Variables

The parameters of primary interest are the joint probabilities $\mathrm{pr}(x = i, y = j)$ or the conditional probabilities $\mathrm{pr}(y = j \mid x = i)$ derived from these. The simple "unadjusted" estimators of these parameters are based on the corresponding sample proportions for the classified variables X and Y and have expectations $\mathrm{pr}(X = i, Y = j)$ under multinomial sampling. Since $\mathrm{Pr}(X = i, Y = j)$ differs in general from $\mathrm{pr}(x = i, y = j)$ the unadjusted estimators are typically biased. Provided the model assumptions (A1)-(A5) hold, the IV estimators of $\mathrm{pr}(x = i, y = j)$ will be asymptotically unbiased although their variances may be larger than those of the unadjusted estimators. The aim of this section is to investigate the extent to which there exists a trade-off in practice between the bias of the unadjusted estimators and the increased variance of the IV estimators. It will be assumed that the model assumptions (A1)-(A5) hold and that the sample is large enough for the IV estimator to be treated as unbiased.

For the numerical investigation in this section we wish to use some "realistic" parameter values. These were determined by rounding the values of estimates for annual flows between the years 1986 and 1987 from analyses in Section 4.2 (reported in Table 3). The values of the five free model parameters not involving W were set to be $K_{21} = 0.03$, $K_{22} = 0.94$, $\mathrm{pr}(x = 2) = \pi = 0.22$, $\mathrm{pr}(y = 2, x = 1) = \theta_1(1 - \pi) = 0.03$ and $\mathrm{pr}(y = 2, x = 2) = \theta_2\pi = 0.19$. Different values of the remaining two free parameters $\varphi_{11} = \mathrm{pr}(W = 1 \mid x = 1)$ and $\varphi_{12} = \mathrm{pr}(W = 1 \mid x = 2)$ are set in the different columns of Table 1. Cramér's V statistic, which measures the association between two binary variables, essentially by scaling the chi-square statistic to a [0,1] interval, is provided as a summary of the strength of association between the variables W and x. For each of the choices of parameter values, Table 1 displays the estimated standard errors of the IV estimators for the PSID sample size $n = 5{,}357$. Table 1 also contains the biases and standard errors of the unadjusted estimator for the same parameter values K_{21}, K_{22}, π, θ_1 and θ_2 and the same sample size.

To illustrate the calculation of the biases of the unadjusted estimators, consider $\mathrm{pr}(x = 1, y = 1)$. The expectation of the unadjusted estimator of this parameter is $\mathrm{pr}(X = 1, Y = 1)$, which is calculated from the given values of K_{21}, K_{22}, π, θ_1 and θ_2 and assumptions (A1)-(A5) as 0.71. This compares with the assumed value of $\mathrm{pr}(x = 1, y = 1)$ of 0.75. The bias is thus $0.71 - 0.75 = -0.04$. The biases of the IV estimators are, as noted above, assumed to be zero. The standard errors of the unadjusted estimators are obtained from standard binomial

Table 1
Biases and Standard Errors under Alternative Hypothetical IVs

			Parameter Values Assumed for IV estimator						
pr($W=1 \mid x=1$)			1.0	0.1	0.1	0.1	0.3	0.1	0.5
pr($W=1 \mid x=2$)			0.0	0.9	0.7	0.5	0.7	0.3	0.3
Cramér's V			1.0	0.74	0.59	0.42	0.34	0.24	0.17
		Standard Errors (× 100)							
Parameter Estimated	Bias (× 100) of Unadjusted Estimator	Unadjusted Estimator	IV Estimator						
pr($x=1, y=1$)	−4.0	0.62	0.68	0.75	0.88	1.13	1.16	1.82	2.05
pr($x=1, y=2$)	3.0	0.32	0.39	0.43	0.51	0.64	0.69	1.03	1.24
pr($x=2, y=1$)	3.0	0.32	0.32	0.37	0.44	0.57	0.66	0.95	1.27
pr($x=2, y=2$)	−2.0	0.51	0.59	0.65	0.73	0.89	1.06	1.42	1.99
pr($y=1 \mid x=1$)	−3.9	0.37	0.50	0.55	0.64	0.81	0.88	1.30	1.58
pr($y=1 \mid x=2$)	12.4	0.60	1.40	1.63	1.95	2.56	2.90	4.30	5.55

Note: 1 = employed, 2 = not employed; n = 5,357; multinomial sampling assumed; biases of IV estimators are zero.

formulae. For example, the standard error of the unadjusted estimator of pr($x=1, y=1$) is $\sqrt{0.71 \times 0.29/5{,}357} = 0.0062$, where 0.71 is the value of $\Pr(X=1, Y=1)$. The standard errors of the IV estimators are obtained from the inverse of the expected information matrix, which is given by $n\sum p_{ijk} H_{ijk}$, where H_{ijk} is the 7×7 matrix of second derivatives of $\log p_{ijk}$ with respect to the seven free parameters. Following differentiation, these parameters are set equal to their assumed values, as indicated above. Note that the standard errors obtained from the multinomial information matrix are likely to be under-estimates because of the complex sampling design employed in the PSID.

There is a clear pattern of the standard errors of the IV estimator increasing as the association between W and x decreases. The amount of increase is fairly similar across all parameters, for example the ratio for V = 0.20 versus V = 1.00 lies between 3 and 4 for all parameters. In all cases the standard error of the IV estimator is greater than that of the unadjusted estimator. The loss of efficiency of the "best" IV estimator (with perfect association between W and x) compared to the adjusted estimator varies between parameters. Roughly speaking, the loss is greater for the conditional parameters than for the unconditional parameters. This loss of efficiency might be interpreted as the effect of adjusting for measurement error in y, which is still necessary even when x is perfectly measured by W. Under this interpretation, the greater relative loss of efficiency for the conditional parameters seems plausible since these are "less dependent" on the parameters of the marginal x distribution which the W information helps to estimate.

To examine the trade-off between the bias of the unadjusted estimator and the increased variance of the IV estimator we have calculated the minimum value of the sample size n necessary for the MSE of the IV estimator to be less than that of the unadjusted estimator. For complex designs the sample sizes should be interpreted as effective sample sizes. Table 2 gives these minimum values under a variety of strengths of association between W and x. If there were no misclassification the entries would all be infinity since the unadjusted estimators would always be more efficient than the IV estimators. For the assumed amount of misclassification given by $K_{21} = 0.03$ and $K_{12} = 0.06$, the sample size required increases rapidly as V decreases. The differences between the rows of Table 2 are partly accounted for by the differences between the rows of Table 1 and partly by differences between the biases of the unadjusted estimator. Thus, the bias of the unadjusted estimator of pr($x=2, y=2$) is relatively small and this leads to the large values in the corresponding row of Table 2. Note that the value of 1 for pr($x=2, y=1$) and Cramér's V = 1 arises because in this case the standard errors of the two estimators are equal (see Table 1) and so the bias of the unadjusted estimators implies that the IV estimator has smaller MSE for any $n \geq 1$.

The main conclusion we wish to draw from Table 2, however, is simply that we may expect there to be a number of practical situations where IV estimation will be worthwhile provided the model assumptions hold, even if the necessary sample sizes are inflated somewhat to allow for complex sampling designs.

4.2 Results for Actual Instrumental Variables

The results in the previous section were based on hypothetical instrumental variables. To provide a more realistic illustration we now consider possible real instrumental variables. The key problem is how to choose a variable W which obeys (A3) and (A4). It seems easier to find a variable which satisfies (A3) than (A4), in particular

Table 2
Sample Size Necessary for MSE of IV Estimator to be less than that of Unadjusted Estimator (Multinomial Sampling)

	Value of Cramér's V assumed for IV estimators						
	1.0	0.74	0.59	0.42	0.34	0.24	0.17
Parameter Estimated	Sample size n required						
$\mathrm{pr}(x=1, y=1)$	28	59	132	300	320	971	1273
$\mathrm{pr}(x=1, y=2)$	31	50	91	184	219	573	843
$\mathrm{pr}(x=2, y=1)$	1	20	51	129	198	476	811
$\mathrm{pr}(x=2, y=2)$	112	227	366	720	1184	2397	5070
$\mathrm{pr}(y=1 \mid x=1)$	42	60	97	183	219	541	818
$\mathrm{pr}(y=1 \mid x=2)$	57	81	121	216	281	633	1061

measured without error obey (A3). However, it seems more difficult to find variables which one is sure are not related to change in employment status and hence obey (A4).

For illustration, we have considered two possibilities. First we have taken W as car ownership ($W = 2$ if the individual owns a car, $W = 1$ if not). This variable is likely to be measured with some error but it seems a reasonable first assumption that this error is unrelated to errors in measuring employment status. For example, in an analysis of errors in recording car ownership in the 1981 British Census, Britton and Birch (1985, p. 67) conclude that "the main problems associated with the small number of discrepancies were those connected with either vehicles out of use or vehicles temporarily available – for example, those hired..." and it seems at least plausible that such errors need have little relation to the kinds of errors in recording employment status. On the other hand, it is plausible that car ownership acts as a proxy for some kind of social or economic status which is related to change in employment status so assumption (A4) seems more questionable. However, for our illustrative purpose we assume (A3) and (A4) hold.

As a second illustration we have taken W to be the lagged employment status in 1985. A problem here is that (A4) effectively implies that individual employment histories follow Markov processes with common transition rates. In fact, transition rates will vary among individuals and this will invalidate assumption (A4) (*e.g.*, van de Pol and Langeheine 1990). Therefore, to allow for departures from assumption (A4), we disaggregated the sample into 16 groups defined by cross-classifying age (4 groups), sex and education (up to college level or not). We then assumed the model held within subgroups and used likelihood ratio tests to assess what parameters were constant across subgroups. These tests only provide a very rough guide since they ignore the complex sampling design of the PSID. There was no significant evidence of differences in the misclassification probabilities K_{ij} across subgroups. Furthermore, within each of the 8 subgroups defined by age × sex there was no significant evidence of differences in $\Pr(W \mid x, \text{subgroup})$ between the 2 education subgroups. Assuming equality of these parameters gave a non-significant likelihood-ratio goodness-of-fit chi-squared value of 52.9 on 46 df (46 is obtained as the number of cells = $16 \times 8 = 128$, less $2K_{ij}$ parameters, less $16 \times 4 = 64$ $\mathrm{pr}(x, y, \text{subgroup})$ parameters, less $8 \times 2 = 16$ $\mathrm{pr}(W \mid x, \text{ subgroup})$ parameters). Combining the parameter estimates for the disaggregated model appropriately gives estimates of the overall flows $\mathrm{pr}(x, y)$.

Table 3 contains estimates of the key parameters for the two choices of instrumental variable and for the disaggregated version of the second choice. We note first that the standard errors for the IV estimator based on car ownership are relatively high. This may be expected from Table 1 since the association between x and W is low (Cramér's V is 0.12). Even so, the resulting adjustments increasing the estimates for the diagonal entries are plausible and the confidence intervals resulting from this IV estimator seem more realistic than those for the unadjusted estimator.

Table 3
Unadjusted and IV Estimates for PSID Data

Parameter	Unadjusted Estimates	IV Estimates		
		IV = Car Ownership	IV = Lagged Employment	IV = Lagged Employment (Disaggregated)
$\mathrm{pr}(x=1, y=1)$	0.719	0.773	0.766	0.757
	(0.006)	(0.033)	(0.008)	(0.007)
$\mathrm{pr}(x=1, y=2)$	0.055	0.011	0.017	0.025
	(0.003)	(0.020)	(0.005)	(0.003)
$\mathrm{pr}(x=2, y=1)$	0.061	0.018	0.024	0.032
	(0.003)	(0.019)	(0.004)	(0.003)
$\mathrm{pr}(x=2, y=2)$	0.166	0.198	0.193	0.186
	(0.005)	(0.027)	(0.007)	(0.006)

Note: Standard errors under multinomial assumptions in parentheses. Disaggregation is by age (4 groups), sex and education (2 groups).

The standard errors for the second choice of instrumental variable are smaller, as expected since the association with X is now higher (Cramér's V is 0.73). Indeed these standard errors are not much larger than those for the unadjusted estimator. The (2 standard error) confidence intervals now do not overlap with the corresponding intervals for the unadjusted estimator for any of the four parameters.

As noted earlier, assumption (A4) is questionable for the lagged employment variable. The disaggregated version of this estimator makes "weaker" assumptions by only requiring (A4) to hold within subgroups. The resulting estimates are seen to be fairly close to the original IV estimator and to have slightly smaller standard errors, perhaps attributable to the use of the additional information on sex, age and education (but see later discussion). It is interesting that the effect of the disaggregation is to diminish the effect of adjustment by a relatively small amount in each case. It seems plausible that departures from (A4) may tend to lead to overadjustment in the IV estimator and that the disaggregation approach here helps to overcome this bias and, for alternative choices of disaggregating variables, enables an assessment of the sensitivity of results to the model specification.

As noted in Section 3 we have often come across IV estimates on the boundary of the interval [0,1]. Of the analyses reported in Table 3 in fact only the disaggregated analysis involved boundary estimates. For the 64 parameters $\mathrm{pr}(x=i, y=j, \text{subgroup})$ for $i, j = 1,2$, subgroup $= 1, \ldots, 16$, five of the estimates were on the boundary (none of the estimates of the remaining 18 parameters, $\mathrm{pr}(W=1 \mid X=1)$ and so forth, were). The standard errors reported in Table 3 treat these parameters as known and hence may underestimate the uncertainty in the estimates of the aggregate $\mathrm{pr}(x=i, y=j)$ parameters.

Table 4
Alternative Estimates of Standard Errors
for Males Aged 26-35 with no College Education

Parameter	IV estimates	Estimated Standard Error	
		Standard	Bootstrap
$\mathrm{pr}(W=1 \mid x=1)$	0.947	0.011	0.011
$\mathrm{pr}(W=1 \mid x=2)$	0.107	0.089	0.091
$\mathrm{pr}(X=1 \mid x=1)$	0.969	0.006	0.007
$\mathrm{pr}(X=1 \mid x=2)$	0.084	0.088	0.075
$\mathrm{pr}(x=1, y=1)$	0.953	0.011	0.012
$\mathrm{pr}(x=1, y=2)$	0	*	*
$\mathrm{pr}(x=2, y=1)$	0.006	0.007	0.006
$\mathrm{pr}(x=2, y=2)$	0.041	0.012	0.011
$\mathrm{pr}(x=1)$	0.953	0.011	0.011
$\mathrm{pr}(y=1 \mid x=1)$	1	*	*
$\mathrm{pr}(y=1 \mid x=2)$	0.128	0.139	0.117

Note: n = 455; "standard" estimators based on observed information matrix, treating parameters estimated at the boundary as known; 10,000 replications of bootstrap; multinomial assumptions.

Table 4 presents alternative estimates of the standard errors for one subgroup, males aged 26-35 with no college education. The estimate of $\mathrm{pr}(x=1, y=2)$ as well as derived estimates, such as $\mathrm{pr}(y=1 \mid x=1)$ lie on the boundary. The "standard" estimates of the standard errors are, as in Table 3, based on the observed information matrix, treating parameters estimated at the boundary as known. Bootstrap standard error estimates (for 10,000 replications) are found to be very close to these standard estimates for parameters with estimates not on the boundary. For the IV estimate of $\mathrm{pr}(x=1, y=2)$ at the boundary no standard estimate of the standard error is available. Indeed it seems to make little sense to estimate the standard deviation of the sampling distribution in this case. It seems more sensible to derive a one-sided confidence interval which may be done either using the profile likelihood method, which gives [0, .016], or using the bootstrap percentile method, which gives [0, .009]. The corresponding intervals for $\mathrm{pr}(y=1 \mid x=1)$ are [.983, 1] and [.990, 1].

5. CONCLUSION

The presence of measurement error can induce substantial bias into standard estimates of transition rates from longitudinal data. If external estimates of misclassification rates are available then a variety of adjustment methods exist. If no such information is available then this paper shows how adjustment for measurement error alternatively can be carried out using instrumental variable estimation.

The main problem, as in conventional instrumental variable estimation, is finding a variable which one can be confident satisfies the conditions required of an instrumental variable. Even if the conditions are satisfied then it is desirable, in order to obtain reasonable precision, that there be a fairly strong association between this variable and the true state. If such a variable can be found then instrumental variable estimation may be useful.

ACKNOWLEDGEMENTS

We are grateful to Wayne Fuller for suggesting the basic idea underlying this paper. Research was supported by grant number H519 25 5005 from the Economic and Social Research Council under its Analysis of Large and Complex Datasets programme.

REFERENCES

ABOWD, J.M., and ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.

ANDERSON, T.W. (1959). Some scaling models and estimation procedures in the latent class model. *Probability and Statistics*, (Ed. U. Grenander). Stockholm: Wiksell and Almquist.

BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis*. London: Griffin.

BIEMER, P.P., GROVES, R.M., LYBERG, L.E., MATHIOWETZ, N.A., and SUDMAN, S. (1991). *Measurement Errors in Surveys* New York: Wiley.

BRITTON, M., and BIRCH, F. (1985). *1981 Census Post-Enumeration Survey*. London: Her Majesty's Stationery Office.

CHUA, T., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

DURBIN, J. (1954). Errors in variables. *Review of the International Statistical Institute*, 22, 23-31.

EDLEFSEN, L.E., and JONES, S.D. (1984). Reference Guide to GAUSS. Applied Technical Systems.

FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*. (Eds. Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S.). New York: Wiley.

FULLER, W.A. (1987). *Measurement Error Models*. New York: Wiley.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage.

HOGUE, C.R., and FLAIM, P.O. (1986). Measuring gross flows in the labor force: an overview of a special conference. *Journal of Business and Economic Statistics*, 41, 111-21.

MADANSKY, A. (1960). Determinental methods in latent class analysis. *Psychometrika*, 25, 183-198.

MARQUIS, K.H., and MOORE, J.C. (1990). Measurement errors in the Survey of Income and Program Participation (SIPP): Program Reports. *Proceedings of the 1990 Annual Research Conference*. US Bureau of the Census, 721-745.

MEYER, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

REIERSOL, D. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9, 1-24.

SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Ch. 3) (Eds. Skinner, C.J., Holt, D., and Smith, T.M.F.). Chichester: Wiley.

SKINNER, C.J., and TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 53, 391-405.

VAN DE POL, F., and DE LEEUW, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15, 118-141.

VAN DE POL, F., and LANGEHEINE, R. (1990). Mixed Markov latent class models. In *Sociological Methodology 1990*, (Ed. C.C. Clogg). Oxford: Basil Blackwell, 213-247.

VAN DE POL, F., LANGEHEINE, R., and DE JONG, W. (1991). PANMARK User Manual. Panel analysis using Markov chains. Version 2.2. Netherlands Central Bureau of Statistics.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 61-71
Statistics Canada

Geographic-Based Oversampling in Demographic Surveys of the United States

JOSEPH WAKSBERG, DAVID JUDKINS and JAMES T. MASSEY[1]

ABSTRACT

Often one of the key objectives of multi-purpose demographic surveys in the U.S. is to produce estimates for small domains of the population such as race, ethnicity, and income. Geographic-based oversampling is one of the techniques often considered for improving the reliability of the small domain statistics using block or block group information from the Bureau of the Census to identify areas where the small domains are concentrated. This paper reviews the issues involved in oversampling geographical areas in conjunction with household screening to improve the precision of small domain estimates. The results from an empirical evaluation of the variance reduction from geographic-based oversampling are given along with an assessment of the robustness of the sampling efficiency over time as information for stratification becomes out of date. The simultaneous oversampling of several small domains is also discussed.

KEY WORDS: Sample design; Stratification; Rare populations.

1. INTRODUCTION

The sponsors of many broad multi-purpose demographic surveys require separate analyses of domains defined by race, ethnicity and income. Equal probability samples generally do not provide sufficient sample sizes for some of these domains to yield the precision needed, making some form of oversampling necessary. This requirement poses interesting methodological problems since there is no registry of the U.S. population from which samples stratified by these domains can be drawn. Housing lists containing identifiers for these domains are maintained at the Bureau of the Census, but they are not available to researchers outside of the Bureau. For surveys requiring face-to-face interviews, outside researchers are thus forced to use area sampling techniques. Even within the Bureau, geography is sometimes used as the basis of oversampling since the lists are only updated once every ten years. This paper describes efficient methods for oversampling the aforementioned domains in the context of area sampling.

Data from the U.S. Decennial Census on concentrations of various demographic domains are publicly available for small geographic units; race and ethnicity are reported for every block and income for every block group. (A "block" is an area bounded on all sides by roads and not transected by any roads. Block groups are combinations of several neighbouring blocks.) These data may be used to inexpensively improve the precision of statistics about rare domains by oversampling blocks or block groups that contain higher than average concentration of members of rare domains and then dropping or subsampling screened persons not in the targeted rare domains. The general theory for this type of sample design was worked out by Kish (1965, Section 4.5). An independent presentation of the theory with examples from the 1960 Decennial Census was given by Waksberg (1973). Further examples and a discussion of alternative methods are given by Kalton and Anderson (1986) and by Kalton writing for the United Nations (1993). In this paper, we extend prior illustrations to cover more domains, update results to 1990, and evaluate empirically the robustness of these methods over time.

We first briefly review the issues involved with screening and subsampling persons not in the targeted domains. Then we review the theory for optimal allocation where the strata are defined in terms of the density of rare populations and apply this theory to several rare populations. The main part of the paper is an empirical evaluation of the reduction in variance reduction from the geographic oversampling of various minority and other rare populations as well as how robust the variance reductions are over time. We also discuss the special problems involved with simultaneous targeting of several rare populations before summarizing our conclusions.

2. SURVEY COST STRUCTURE AND THE SCREENING DECISION

Let U stand for some target universe such as persons or households for which a sampling frame exists. Let D stand for some small domain of particular interest such as black persons that cannot be separately identified from the balance of U at the time of sampling. Let $\boldsymbol{Y}$ be a vector of characteristics of interest such as annual income, employment status, and number of doctors' visits in the last year. In some surveys, the only objective is estimation of the distribution of $\boldsymbol{Y}$ on D. In such surveys, members of U-D that are discovered in the course of screening sampled members of U will be dropped from the sample. A general inexpensive interview

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, formerly of Westat Inc., now deceased.

questionnaire is used for the screening to determine who is eligible for a full questionnaire.

In other surveys, estimation of the distribution of Y on D and on U are both important objectives. For such a survey, at least some of the members of U-D that are discovered in the course of screening interviews will be retained for full interviews. If geographic-based oversampling is used, the initial sample will contain an oversample of those members of U-D who happen to reside in areas with heavy concentrations of D. Even when U-D is of interest, this oversampling of U-D in areas with high concentrations of D is usually undesirable since resulting variation in probabilities of selection for U-D leads to unnecessarily large design effects for statistics both about U and about U-D. These larger design effects mean that the extra sample size for U-D will usually result in only a trivial decrease in variances for statistics about U-D. Generally, the funds expended on the extra interviews with U-D would be better spent on increasing the total initial sample size.

It is fairly easy to set up subsampling procedures that result in an equi-probability sample of U-D. The subsampling can be done centrally after the completion of the entire screening operation, or it can be done by the interviewer while still in the sample household after obtaining data on household composition. Techniques have been developed that make the subsampling process very easy for the interviewer (Waksberg and Mohadjer 1991). Interviewers do not need to be trained to carry out random draws. With paper and pencil survey instruments, interviewers are given house-by-house pre-interview instructions about which domains can be interviewed at which households. These instructions are randomized centrally prior to screening to yield the desired sampling rates. Alternatively, with CAPI, the subsampling can be programmed and carried out automatically in the laptop computer used for CAPI; the computer notifies the interviewer which households are to be retained for the full interview and which ones to reject as a result of subsampling.

Whether it is better to keep all sampled members of U-D or to subsample them depends on the relative sizes of U and U-D, the precision requirements for both and on the relative costs of full interviews and the shorter screening interviews. Let $c_\star$ be the variable cost associated with sampling a single member of U and collecting and processing all data of interest about that member. Let c' be the variable cost associated with sampling, screening, and then dropping a single member of U. Let $c = c_\star/c'$, be the ratio of the cost of a full interview to the cost of a screening interview. If c is much greater than 1, then subsampling should be considered for the survey that has interest in U-D even though subsampling of U-D will introduce some additional complexity into survey operations. Given that the full interview is by definition longer that the screening interview, it should always be the case that c is at least slightly greater than 1. On panel and longitudinal surveys, the cost of all follow-back interviews should be counted as part of $c_\star$, typically making the cost of a full interview many times larger than the cost of a screening interview; *i.e.*, $c >> 1$. The same will be true of surveys that involve the collection of physical specima requiring expensive laboratory work and of surveys that require expensive experts (such as medical doctors) to participate in the primary data collection. For such surveys, we would highly recommend that geographic-based oversampling not be employed by itself, but rather, in conjunction with screening and subsampling. For a door-to-door survey with a single interview by a standard grade interviewer (trained to ask questions and record answers but not to make any technical or anthropological assessments), c is frequently in the range of 3 to 5. This is large enough in many applications to justify the complication of subsampling U-D in oversampled areas.

3. FORMING THE STRATA

We assume that even though D cannot be separated from U at the time of sampling, there is some information available about the distribution of D and U across a set of geographically defined entities. In the United States, the natural entities are blocks or block groups (BGs) and information for these entities is supplied by the decennial census. (Prior to the 1990 decennial census, blocks were not defined in rural areas; larger entities called "enumeration districts" were used for oversampling.) The U.S. Bureau of the Census makes data on the racial and ethnic composition of blocks publicly available along with mapping information so that these blocks can be identified years later by any survey organization. Income data are only made available at the BG level.

Standard practice calls for the stratification of the blocks or BGs by the local concentration of D. Thus, all blocks where D constitutes less that 10 percent of the block's total population might constitute one stratum. Further cutpoints for defining the strata might be 30 percent, and 60 percent, yielding a total of four strata. There has been little empirical study of the optimal number of strata nor of the optimal cutpoints. In general, more strata will yield more efficient designs, but, at some point, the operational complexities of a large number of strata outweigh the gains in efficiency. Conventional wisdom dating back to Kish (1965) holds that a fairly small number of strata will achieve most of the gains attainable through stratification.

4. OPTIMAL ALLOCATION FOR A SINGLE DOMAIN

Our objective is to adapt the general formulas for optimum allocation of a stratified sample to apply to the reduction in variance due to geographic-based oversampling. The derivations are essentially those given by Kish (1965) using the notation of Kalton in United Nations (1993). Let the population be divided into a number of strata as discussed above. Let N be the size of the total population and N_h be the

size of the total population within the h-th stratum. Let P_h be the proportion of the h-th stratum that consists of members of D. Let P be the overall proportion of the population that belongs to D. We may use the prior decennial census to estimate P_h and P, or we may use some more recent large survey that carried block and/or BG codes for every sample household/person so that matching to the last decennial census will yield the stratum identification for every sample household/person.

We assume that c is constant across the strata even though this may sometimes not be very accurate. For example, interviewing in blocks with high concentrations of American Indians, Eskimos or Aleuts almost always means interviewing in remote locations with difficult transportation issues. However, estimation of even a national average for c is difficult for most survey operations. It will not generally be possible to get estimates by stratum.

We also assume that the distribution of Y on D is constant across the strata. More specifically, we assume that

$$E(Y \mid D \text{ and } h) \equiv E(Y \mid D) \quad \text{and that}$$

$$\operatorname{Var}(Y \mid D \text{ and } h) \equiv \operatorname{Var}(Y \mid D),$$

where the expected value and variance are with respect to the population, not the sample design. This is usually not a very good assumption, but given a vector of characteristics of interest, the components of the vector will usually behave differently across the strata so there is no point in trying to be more exact. Lastly, we assume that the sampling fractions are small enough in all the strata to make the finite population correction factors ignorable.

Given these assumptions, the optimal sampling fraction for the h-th stratum for a survey where all screened members of U-D are dropped is

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \tag{1}$$

where k is a constant determined by either precision requirements or budget constraints. (For a proof of (1), see either of the sources referenced above. This allocation rule is an application of Neyman allocation.) If $c=1$, (*i.e.*, screening is as expensive as interviewing), then this proportionality reduces to $f_h \propto \sqrt{P_h}$, which can yield allocations quite different from an equi-probability sample across strata. However, if the cost of screening is far less than the cost of interviewing (*i.e.*, $c >> 1$) and D is not extremely rare (*i.e.*, P_h is not close to zero), then this relationship results in close to a flat set of sampling intervals, which is equivalent to allocation in proportion to total population.

Given a fixed budget of B, k is determined by the cost equation

$$B = \sum_h N_h f_h c' \left[P_h c + (1-P_h)\right]. \tag{2}$$

To obtain a simple random sample of size n from domain D would require selecting a screening sample of size n/P, resulting in a total cost of

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \tag{3}$$

By equating these two costs, we can solve for the constant of proportionality in (1) and get:

$$k = \frac{n\left(c-1+\frac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c-1+\frac{1}{P_h}}}. \tag{4}$$

To calculate the benefits of this allocation realistically, it is necessary to acknowledge the fact that the estimates of P_h that are used to guide the allocation will be somewhat out of date by the time that the survey is actually conducted. Let A_h be the proportion of D actually to be found within the h-th stratum at the time of sampling and data collection. It is assumed that P is unchanged even though the distribution across strata changes according to A_h. By letting $NP = N_D$ and $N_D A_h = N_{Dh}$ it can readily be shown that the actual sample size, n_D, that will be achieved on D is given by

$$n_D = \sum_h NPA_h f_h. \tag{5}$$

From Kish (1965), this sample will have higher variance than a simple random sample of the same size on D. The variance inflation factor or design effect associated with the differential sampling rates across strata is the well-known

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h / f_h\right). \tag{6}$$

Thus, the *effective* sample size associated with the geographic-based oversampling is

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h / f_h\right)}. \tag{7}$$

Substitution of formulae (1) and (4) into (7) yields

$$\frac{n_D}{deff} = \frac{n\left(c-1+\frac{1}{P}\right)}{\left(\sum_h A_h \sqrt{c-1+\frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP}\sqrt{c-1+\frac{1}{P_h}}\right)}. \tag{8}$$

This formula allows us to compare the variance for an arbitrary statistic on domain D given geographic-based oversampling with the variance for the same statistic given a simple random sample of D of the same total cost B. Formula (8) can be rewritten algebraically such that the proportion of simple random sample variance that is eliminated by the geographic-based oversampling is given by

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} = 1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \quad (9)$$

It is definitely possible for this reduction to be negative, meaning that a simple random sample would have provided lower variance for the same cost. This is most likely to happen when there exists a stratum for which $NPA_h >> N_h P_h$, meaning that there exists a stratum which was thought to have a very small portion of D but, in fact, has quite a significant portion of D. Note that if $P_h \equiv P$, then no variance reduction can be expected from geographic-based oversampling. Also, as c goes to infinity for fixed P (equivalent to screening becoming cheaper and cheaper relative to full interviews), the variance reduction approaches zero. Given the extra complication of a stratified sample, this means that for large c and moderate P, the sample designer should consider drawing a simple random sample instead of a stratified sample. Geographic-based oversampling increases in value as P approaches zero, c approaches 1, and D becomes more concentrated in a single stratum. As the small domain of interest, D, becomes more concentrated in a single stratum the sample becomes more efficient, since there are fewer cases from D in the remaining strata with large differential. The potential reductions in variance due to geographic-based oversampling under a number of conditions are shown empirically for several demographic domains in the section below.

5. EMPIRICAL EVALUATION

Equation (9) is quite difficult to evaluate for domains of interest. Data on P_h can be obtained from summary tapes from the decennial censuses that are published at the block, block group, and enumeration district levels by the Bureau of the Census. This allows one to define reasonable strata and to evaluate equations (1) through (4). If one were to assume that the P_h are static over time, then the rest of the equations could also be evaluated. However, Americans tend to move frequently, and the racial and ethnic composition of many blocks change in that process (Judkins, Massey and Waksberg 1992). To the extent that members of D move into areas where they were previously not common, the benefits of the geographic-based oversampling diminish. Not wishing to overstate the benefits of the procedure, we searched for some method to get reasonable estimates of the A_h at postcensal time points. Matching block- or BG-level data for two consecutive censuses might appear to be a good solution but is not possible. Up to now, blocks have been defined and labelled independently from census to census with no attempt to preserve definitions for longitudinal. Thus, alternate information sources are required to estimate A_h.

For the analysis of the benefits of geographic-based oversampling for the black and Hispanic populations, micro-level data from current household surveys conducted by the Census Bureau turned out to be a good source of information on the A_h. Specifically, we used data from the 1988 National Health Interview Survey (NHIS). Staff at the Census Bureau prepared a special tape for us that gave the 1980 block group or enumeration district code for almost all households interviewed in the 1988 NHIS in residences built prior to 1980. (Residences constructed during the 1980s would have been sampled for the NHIS from building permits rather than by area sampling. Due to technical difficulties, block and block group labels are not attached to such sample dwellings.) We then matched the 1988 NHIS against 1980 Census summary files by block group or enumeration district in order to classify NHIS households into strata defined by concentrations of blacks and Hispanics in 1980. Using survey weights, we were then able to estimate the distribution of various domains across those strata. (Housing built during the 1980s was assumed to be in the stratum with the lowest concentration of the rare domains.) Similar operations could have been carried out for Asians, Pacific Islanders, American Indians, Eskimos, Aleuts, and persons with low income but were not.

Tables and charts in the balance of the paper will refer to data at several points in time and from several sources. It is useful to bear in mind that the data used to form the strata do not have to be the same as the data used to allocate the sample, and that the data used to evaluate the sample may be from a third point in time or source. We have the following combinations in this paper:

Label	Source of stratification data	Source of allocation data	Source of evaluation data
80/80/80 BG	1980 Census (BG level)	1980 Census	1980 Census
80/80/88 BG	1980 Census (BG level)	1980 Census	1988 NHIS
80/88/88 BG	1980 Census (BG level)	1988 NHIS	1988 NHIS
90/90/90 BG	1990 Census (BG level)	1990 Census	1990 Census
90/90/90 blk	1990 Census (block level)	1990 Census	1990 Census

Table 1
Residential Clustering of Blacks

Density stratum (Blacks as a percent of the stratification unit in the year of stratification)	Percentage of blacks living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Total populations (1000s)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Blacks as percent of nation in measurement year	11.7	12.0	12.1	12.1				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

6. OVERSAMPLING THE BLACK POPULATION

Table 1 shows various aspects of residential segregation for the black population in the U.S. that are important to know about when designing a population survey. Although the percentage of blacks living in densely black (60+ percent) block groups declined between 1980 and 1990, it is clear that blacks were still strongly segregated. The columns about the population in 1988 are particularly important since they show the dynamics of the stratification data over time. By 1988, the percentage of the black population living in the block groups that were less than 10 percent black in 1980 had doubled, from just 9.7 percent of blacks to 20.5 percent. This has major implications for the efficacy of geographic-based oversampling as will be shown below. It is also interesting to note that the total population in the block groups that were densely black (*i.e.*, over 60% black) in 1980 actually declined by about 2 million persons between 1980 and 1988. At least part of this shift came from abandonment of some old housing and neighbourhoods. Concentration levels are sharper at the block level than at the block group level in 1990, as would be expected. (Block level data are not available for the whole nation from 1980.) Although sampling blocks is slightly more costly than sampling block groups (due to the larger number of blocks and the need to make provisions for blocks that have fewer inhabitants than the desired sample cluster size), it does allow sharper focus on the targeted domain.

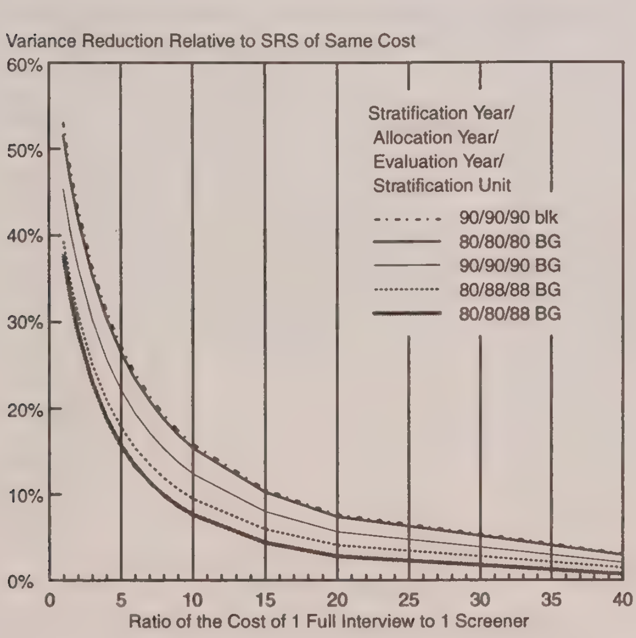


Figure 1. Variance Reduction from Geographic-based Oversampling for Blacks

Figure 1 summarizes the implications of the density data shown in Table 1 for oversampling blacks. This figure shows the substantial effect of c on the efficiency of geographic-based oversampling. For values of c beyond 20, the best way to sample the black population is probably just to screen an equi-probability sample.

The figure also illustrates the danger of relying upon the stratification data to evaluate the benefits of geographic-based oversampling. The 80/80/80 line shows the variance reductions that could be made if there were no change over time in the distribution of the black population across the density strata defined in terms of 1980 block group data. The 80/80/88 line shows the actual variance reductions that are possible in 1988 for the same strata and allocation. At $c = 5$, the variance reduction given a static distribution is 26 percent, while the variance reduction given observed changes in the distribution is just 16 percent. We examined whether allocating the sample across the old strata according to new distribution data could improve the actual variance reduction in 1988. The answer is yes, but not by much. The 80/88/88 shows the variance reductions that are possible using the 1988

distribution across the 1980 strata to guide the allocation for a survey conducted in 1988. At $c = 5$, the variance reduction given this allocation is 18 percent, a very modest improvement over the 16 percent variance reduction possible with the allocation guided by the old distribution. This led us to conclude that the major problem was the old stratification itself. By 1988, the extent of migration by the black population from block groups that were densely black in 1980 into block groups that had lower concentrations of black populations in 1980 was so great as to cut the variance reduction achievable through oversampling almost in half. The shift of the black population into block groups with lower concentrations of blacks in 1980 results in more sample blacks with large weights thus increasing the variability among weights which increases the variance. Nonetheless, the variance reductions indicated by the 80/80/88 line for $c < 10$ are certainly large enough to be useful.

Turning attention to the 1990 data in Figure 1, we observe that the 90/90/90 BG line is consistently several points below the 80/80/80 line, indicating that geographic oversampling at the block group level is likely to be slightly less useful during the 1990s than it was during the 1980s. This is a reflection of the slight reduction in segregation of the American black population in 1990 compared to 1980 noted above. On the other hand, the 90/90/90 blk line is almost exactly the same as the 80/80/80 line, indicating that the geographic oversampling at the block level can be expected to be as effective during the 1990s as it was at the block group level in the 1980s. Although data have not yet been collected on the distribution of the black population in the late 1990s across 1990 density strata, we would expect that migration has continued and that therefore the gains indicated by the 1990 lines should probably be reduced (along the general trend indicated by the 80/80/88 line) when projecting savings into the late 1990s and the first few years after 2000.

7. OVERSAMPLING HISPANICS

Table 2 shows various aspects of residential segregation for Hispanics in the U.S. that are important to know about when designing a population survey. Several points are interesting to note. First, it appears that Hispanics (unlike blacks) became slightly more segregated between 1980 and 1990. Other patterns, however, are similar for the black and Hispanic populations. In 1980, 30 percent of the Hispanic population lived in block groups that were 60 percent or more Hispanic. By 1988 these same block groups contained only

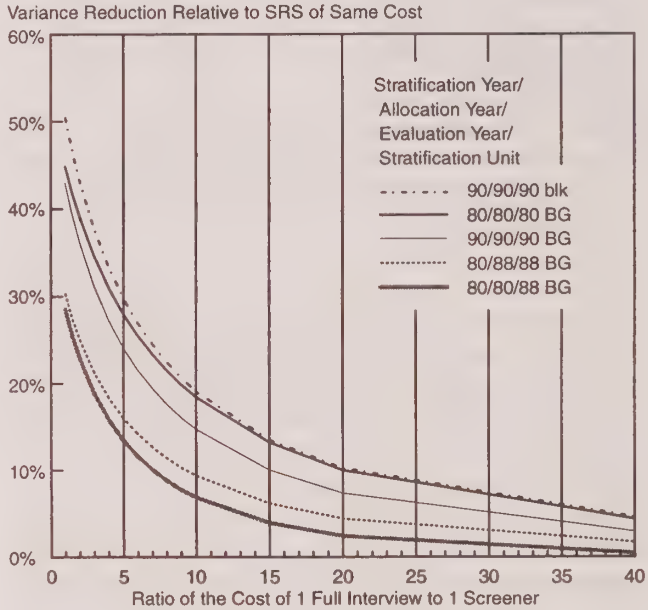


Figure 2. Variance Reduction from Geographic-based Oversampling for Hispanics

Table 2
Residential Clustering of Hispanics

Density stratum (Hispanics as a percent of the stratification unit in the year of stratification)	Percentage of Hispanics living in the stratum in the indicated year				Percentage of the total population living in the stratum in the indicated year			
Measurement year	1980	1988	1990	1990	1980	1988	1990	1990
Stratification year	1980	1980	1990	1990	1980	1980	1990	1990
Stratification unit	BG/ED	BG/ED	BG	Block	BG/ED	BG/ED	BG	Block
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Total populations (1000s)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Hispanics as percent of nation in measurement year	6.4	8.1	9.0	9.0				

Sources: 1980 Decennial Census (Westat tabulation)
1988 National Health Interview Survey (Westat tabulation)
1990 Decennial Census (Westat tabulation)

about 21 percent of the Hispanic population. In contrast, the percent of Hispanic population living in the 1980 block groups that were less than 5 percent Hispanic increased from 15 percent in 1980 to 29 percent in 1988. These changes reflect both a shift of the Hispanic between areas and the increase in the Hispanic population coming into the United States. The restratification of the Hispanic population using 1990 data shows patterns similar to the 1980 distribution patterns.

Figure 2 summarizes the implications of these segregation data on oversampling schemes. The curves show the same general patterns as the black curves. Geographic-based oversampling appears to be a useful tool for values of $c < 10$. Again though, it is important to be mindful of the effect of migration on the variance reduction. The gap between the 80/80/80 and 80/80/88 lines is greater for Hispanics than for blacks, particularly for $c < 5$. At present, we do not have a good basis for predicting whether this will be as true in the 1990s as it was in the 1980s.

8. OVERSAMPLING OTHER RACIAL MINORITIES

Tables 3 and 4 show segregation data for Asians and Pacific Islanders and for American Indians, Eskimos and Aleuts, respectively. Figures 3 and 4 show corresponding implications for oversampling these domains. Data from 1980 and 1988 were not tabulated for this work because the 1990 data are not encouraging for the inexpensive oversampling of these populations even with the use of stratification by density. The percent reductions in variance are quite large, greater than those for the black and Hispanic populations, since the amount of screening that would otherwise be required is much larger. However, the rarity of these populations in the U.S. means that very large screening samples are still required in order to get respectable interviewed sample sizes. For example, with a cost ratio of 3, even with geographic-based oversampling, it is necessary to screen 61,000 persons (or about 24,000 households) in order

Table 3
Residential Clustering of Asians and Pacific Islanders

Density stratum (Asians and Pacific Islanders as a percent of the 1990 block or block group in 1990)	Percentage of Asians and Pacific Islanders living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Total population (1000s)	6,968	6,968	248,710	248,710
Asians and Pacific Islanders as percent of nation in measurement year	2.8	2.8		

Sources: 1990 Decennial Census (Westat tabulation)

Table 4
Residential Clustering of American Indians, Eskimos and Aleuts

Density stratum (American Indians, Eskimos and Aleuts as a percent of the 1990 block or block group in 1990)	Percentage of American Indians, Eskimos and Aleuts living in the stratum in 1990		Percentage of the total population living in the stratum in 1990	
Stratification unit:	BG	Block	BG	Block
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Total population (1000s)	1,793	1,793	248,710	248,710
American Indians, Eskimos and Aleuts as percent of nation in measurement year	0.7	0.7		

Sources: 1990 Decennial Census (Westat tabulation)

to obtain a sample of American Indians, Eskimos and Aleuts with precision equal to a (theoretical) simple random sample of 1,000 persons from this domain. (Of course, to successfully screen 24,000 households, more housing units would have to be selected to allow for vacants and nonresponse). The comparable number for Asians and Pacific Islanders is 18,000 persons or roughly 7,000 households.

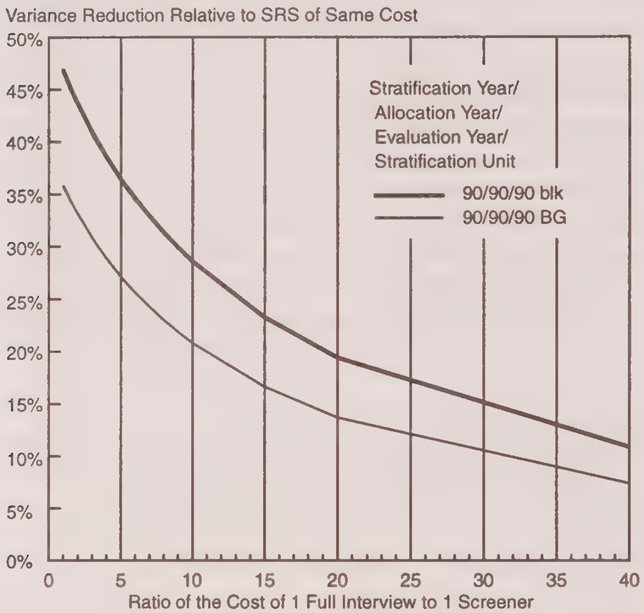


Figure 3. Variance Reduction from Geographic-based Oversampling for Asians and Pacific Islanders

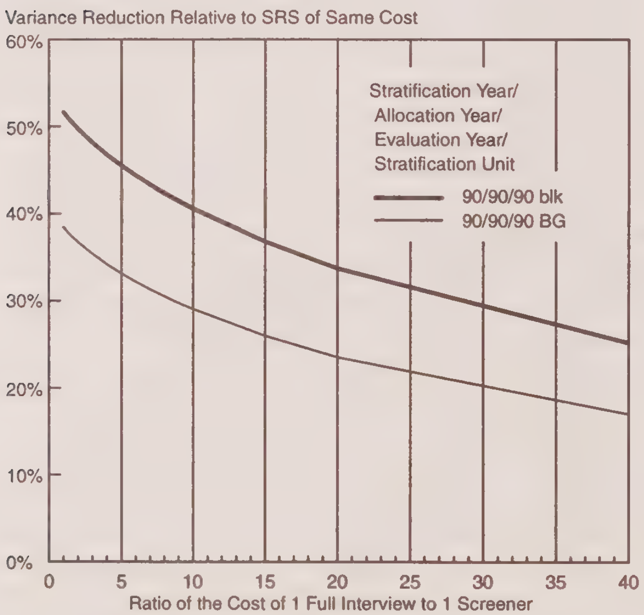


Figure 4. Variance Reduction from Geographic-based Oversampling for American Indians, Eskimos and Aleuts

9. OVERSAMPLING THE POOR

Table 5 shows the 1990 distribution of the low income population by block groups classified according to the proportion of low-income population in the BG. The BGs in each of the classes depends on the definition of low income. The figures shown in the table are the percentages of low-income persons in each class. Table 5 shows a rather flat distribution of low income among the classes for all three definitions in 1990. Data (not shown) from the 1970 decennial census and the Current Population Survey indicate that segregation of persons below the poverty level increased between 1970 and 1990 (Waksberg 1995), but the segregation is still far less than the segregation of racial and ethnic groups. The concentrations are somewhat greater for persons under 150 percent than for the other two definitions but, even for this group, it is considerably less than for racial and ethnic groups. As can be seen, with this definition, only about 25 percent of the poor live in BGs where 50 percent or more of the population is poor. The comparable percentages are 19 percent for persons below 125 percent of poverty and only 13 percent for persons below 100 percent of poverty. Such distributions imply that oversampling households in the strata with relatively high percentages of low-income persons will not be much better than oversampling and screening the entire sampling frame unless the full interview costs are only slightly higher than screening costs.

Figure 5 shows the ratio of the variance of the optimum sample to an SRS at the same cost, for statistics relating to the low-income populations. Interestingly, despite the greater concentration associated with the broadest definition of low

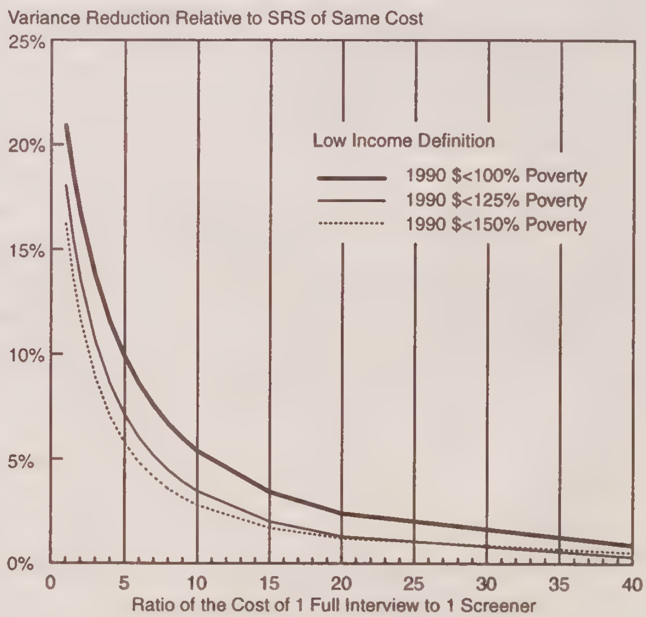


Figure 5. Variance Reduction from Geographic-based Oversampling for Persons with Low Income

Table 5
Residential Clustering of the Low Income Population

Density stratum (Persons with low income as a percent of 1990 block group in 1990 according to various definitions of low income)	Percentage of persons with low income living in the stratum in 1990			Percentage of the total population living in the stratum in 1990		
Low income definition:	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty	$ < Poverty	$ < 125% of Poverty	$ < 150% of Poverty
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21.0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Total populations (1000s)	31,797	42,316	52,521	248,710	248,710	248,710
Persons with low income as percent of nation in measurement year	12.8	17.0	21.1			

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

income, the reduction in variance for geographic-based oversampling is strongest for the narrowest definition because it requires more screening and thus has more to gain from a sampling strategy that reduces screening. For all three definitions, there appear to be moderate advantages to oversampling when c is under 3 or 4, about a 10 or 15 percent reduction in variances. When c is as large as 10, the gains are very slight, and there is virtually no advantage to oversampling BGs with high levels of poverty when c is 20 or larger. Of course, migration must be taken into account here as well, but we did not obtain the necessary data. Due to the effects of migration, the actual variance reductions will almost certainly be smaller than those shown in the chart. Furthermore, the income data in the 1990 Census are based on a one-sixth sample. The sample size in a typical block group was a little under 100 households. The classification of blocks according to percentage of low-income persons therefore has a fair amount of fuzziness to it, and many block groups will not be in the categories that Census data assign them, but in neighbouring classes, further weakening the variance reductions that can be achieved with geographic-based oversampling. As a result of these factors, it is unlikely that geographic-based oversampling will improve the efficiency. In fact, by mid-decade or later, it may actually result in an increase in variance. A related unpublished study by Waksberg in 1989 showed similar results when considering the possibility of merging ZIP-code level summary income data onto banks of telephone numbers used in RDD sampling. The gains achievable through stratification appear quite limited.

An examination of more detailed tables (not shown) indicates that the effectiveness is about the same for various types of geographic breakdowns, *e.g.*, states, large or small MSAs, central cities, suburban areas, and nonmetropolitan areas. Conclusions drawn from this analysis will thus approximately apply to subnational surveys.

However, geographic-based oversampling is an extremely effective tool for the low-income black and Hispanic populations. As shown in Table 6, blacks and Hispanics living in poverty are highly concentrated and others living in poverty are not. The left-hand side of Table 6 indicates the distribution of the poor black, Hispanic, and other populations across density strata defined in terms of poverty rates specific to the domain of interest. Interpreting one example from the left side, 32 percent of poor Hispanics lived in 1990 in block groups where the poverty rate for Hispanics was over 50 percent. The right hand side indicates the distribution of the poor black and Hispanic populations across density strata defined just in terms of the local concentrations of blacks or Hispanics without regard to income levels. Interpreting one example from the right side, 44.8 percent of poor Hispanics lived in 1990 in block groups where Hispanics constituted over 60 percent of the local population. From these numbers, we infer that over 90 percent of both poor blacks and poor Hispanics live in areas with above average concentrations of their respective racial/ethnic groups. This means that a sampling strategy that oversamples blocks with high black or Hispanic concentrations will automatically yield disproportionately large numbers of poor blacks and Hispanics. Furthermore, almost no poor blacks or poor Hispanics live in areas with low poverty rates for their groups. This stands in marked contrast to the patterns for poor people who are neither black nor Hispanic. It appears that many poor nonhispanic whites live in close proximity to more well-off whites, possibly because poverty tends to be a transitory phenomenon for them, or perhaps because they are retired and purchased their homes when they were in better circumstances.

Table 6
Residential Clustering of the Low Income Population by Race and Ethnicity

Density stratum (Poverty rate in 1990 for persons of the indicated race/ethnicity within the block group in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990			Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of persons with the indicated race/ethnicity and income below the poverty line living in the stratum in 1990		
	Domain				Domain		
	Blacks	Hispanics	Others		Blacks	Hispanics	Others
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Total populations (1000s)	8,557	5,536	17,975	Total populations (1000s)	8,557	5,536	17,975

Sources: 1990 Decennial Census (Westat tabulation of STF-3)

10. SIMULTANEOUS OVERSAMPLING OF SEVERAL RACE-ETHNIC DOMAINS

In general, geographic-based oversampling can be used as easily and effectively for targeting multiple race-ethnic domains as for a single race-ethnic domain. In fact, the optimal sampling rates for the strata with high concentrations of each of the targeted domains will be about the same as if only it were being targeted. However, the overall level of screening will be increased since the number of areas with high sampling rates will increase with the number of targeted domains. Both these observations are due to the limited overlap between the highly segregated areas of the examined racial and ethnic minorities.

Table 7 presents some data on this subject from the 1990 Decennial Census. The only domains that overlap significantly in their concentrated areas are Hispanics and Asians and Pacific Islanders, and even that overlap only works one way. Since there are so many more Hispanics in the U.S. than Asians and Pacific Islanders, the proportion of Hispanics that live in blocks with Asian /Pacific Islander populations over 10 percent of the local population is only 13.7 percent while the percent of Asians and Pacific Islanders that live in blocks with Hispanic populations over 10 percent of the local population is a high 40.8 percent. The practical significance of this particular overlap is probably slight, however, since it would take such a large screening sample (both in and out of highly concentrated areas) to find enough Asians and Pacific Islanders to meet moderate precision requirements that such

Table 7
Residential Mixing of Minorities

Density stratum (Indicated minority as a percent of 1990 block in 1990)	Percentage of blacks living in the stratum in 1990			Percentage of Hispanics living in the stratum in 1990			Percentage of Asians and Pacific Islanders living in 1990			Percentage of American Indians, Eskimos and Aleuts living in 1990		
	Stratification domain			Stratification domain			Stratification domain			Stratification domain		
	Hispanic	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Asian and Pacific Islander	American Indian, Eskimo and Aleut	Black	Hispanic	American Indian, Eskimo and Aleut	Black	Hispanic	Asian and Pacific Islander
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: 1990 Decennial Census (Westat tabulation)

a screening sample would probably find enough Hispanics without resorting to disproportionate allocation of the sample to blocks with higher concentrations of Hispanics.

11. CONCLUSIONS

For household surveys in the U.S., geographic-based oversampling using data from the most recent decennial census is a useful sampling strategy for improving the precision of statistics about the black and Hispanic populations provided that the cost of full interviews is less than 5 to 10 times the cost of screener interviews. It is also a useful strategy for improving the precision of statistics about the Asian/Pacific Islander and American Indian/Eskimo/Aleut populations, even at very high ratios of the cost of full interviews to the cost of screener interviews.

However, this does not mean that a survey of reasonable cost can be designed to simultaneously provide highly precise statistics about all these domains while maintaining desired precision levels for the total population. Most demographic surveys require reasonable precision for both targeted domains and for the total population. Shifting some portion of the full interviews from the white nonhispanic population to the other domains is bound to decrease the precision of statistics about the total population. It is generally useful to strike a balance between precision attained for subpopulations and the total population. The point of this observation is merely that geographic-based oversampling does not obviate the need to select very large samples and conduct many screening interviews when trying to obtain precise statistics about rare domains at the lowest possible cost. Furthermore, precise statistics about rare domains will continue to be expensive even when using geographic-based oversampling.

For surveys of low-income persons, only small gains are possible with geographic-based oversampling, and those only when the cost of a full interview is only a few times larger than the cost of screening and dropping a household. Most of these gains are likely to disappear when deterioration over time is taken into account. In fact, by the middle of a decade or later, when Census data become seriously outdated, there is the distinct possibility that geographic-based oversampling could reduce efficiency rather than improve it because of migration of the poor and sampling error in measuring poverty at the block group level. Geographic-based oversampling is a useful tool, however, when the focus of interest is on the black or Hispanic poor.

ACKNOWLEDGMENTS

This research was conducted by Westat Inc. under contract 200-89-7021 sponsored by the National Center for Health Statistics, Centers for Disease Control and Prevention. David Judkins and James Massey participated in the project while they were with Westat and NCHS, respectively. The authors would like to gratefully acknowledge the programming contributions of John Edmonds and Robert Dymowski of Westat and to thank the referees for their useful comments and suggestions on an earlier version of the paper.

REFERENCES

JUDKINS, D., MASSEY, J., and WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association,* 51-60.

KALTON, G., and ANDERSON, D.W. (1986) Sampling rare populations. *Journal of the Royal Statistical Society, Series A*, 149, 1, 65-82.

KISH, L. (1965). *Survey Sampling.* New York: Wiley.

MASSEY, J., JUDKINS, D., and WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 75-84.

UNITED NATIONS (1993). Sampling Rare and Elusive Populations. Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 497-502.

WAKSBERG, J., and MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 350-355.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 73-75
Statistics Canada

A Modified Random Groups Standard Error Estimator

WILLARD C. LOSINGER[1]

ABSTRACT

The standard error estimation method used for sample data in the U.S. Decennial Census from 1970 through 1990 yielded irregular results. For example, the method gave different standard error estimates for the "yes" and "no" response for the same binomial variable, when both standard error estimates should have been the same. If most respondents answered a binomial variable one way and a few answered the other way, the standard error estimate was much higher for the response with the most respondents. In addition, when 100 percent of respondents answered a question the same way, the standard error of this estimate was not zero, but was still quite high. Reporting average design effects which were weighted by the number of respondents that reported particular characteristics magnified the problem. An alternative to the random groups standard error estimate used in the U.S. census is suggested here.

KEY WORDS: Census; Variance estimation; Random groups; Design effect.

1. INTRODUCTION

During the 1990 Decennial Census, all respondents were asked to provide information on certain data items (called 100-percent data). Most respondents provided this information on the census short form. In addition, a systematic sample (ranging from one-eighth to one-half, but averaging about one-sixth) of respondents provided information for more data items (sample data) on the census long form.

Rather than providing standard error estimates for each published sample data estimate, the Census Bureau published tables of generalized design effects. For any sample data estimate, data users were instructed to create a standard error assuming simple random sampling (either using the standard formula or from a table) and a one-in-six sampling rate. Then, data users were to multiply this standard error by a generalized design effect (provided in another table). The table of generalized design effects listed design effects by data item type and percent of persons or housing units included in the sample (Table 1 provides the design effects published for 1990 U.S. census sample data for Vermont). For example, for all published sample estimates that dealt with occupation, a data user would find four generalized design effects for occupation: one for each of four sampling rate categories for persons in the report. To estimate the standard error for the number of teachers in a published report, a data user would multiply the simple-random-sampling standard error (assuming a one-in-six sampling rate, derived from the formula or table of standard errors) by the design effect for occupation data items for the reported sampling rate. The data user could then use the estimated number of teachers and standard error to construct a confidence interval. More details on the use of the table of design effects are available in the Accuracy of the Data section for any sample data product (U.S. Bureau of the Census 1993, for example).

2. ESTIMATION OF STANDARD ERRORS

A random-groups approach was used to estimate standard errors for the census sample data. The United States was divided into just over 60,000 distinct areas (called weighting areas--areas for which sample weights were derived). For each weighting area, sample units (a sample unit being either a housing unit or a person residing in a group quarter) were assigned systematically among 25 random groups. Thus, it was thought that each random group so formed met the requirement of having approximately the same sampling design as the parent sample (Wolter 1985).

For each of the 25 random groups, a separate estimate of the total for each of 1,804 sample data items was computed by multiplying the weighted count for the sample data item within the random group by 25. For each data item for which the total number of people with a particular characteristic was estimated from the sample data, the random-groups standard error estimate was then computed from the 25 different estimates of the total from the random groups:

$$S_{\mathrm{RG}} = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{Y}_i - \hat{Y})^2}{24}}$$

where n represents the unweighted number of persons in the sample within the weighting area; N represents the census count of persons within the weighting area; $\hat{Y}_i$ represents the estimate of the total for the data item achieved by multiplying the weighted count for the data item within the i-th random group by 25; and $\hat{Y}$ is the weighted count for the data item (*i.e.*, the sample estimate) within the weighting area.

[1] Willard C. Losinger, U.S. Department of Agriculture: APHIS:VS, CEAH, 555 South Howes Street, Suite 200, Fort Collins, CO 80521, U.S.A.

Table 1
Design Effects Published for 1990 U.S. Census Sample Data for Vermont

Characteristic	Percent of persons or housing units in sample			
	< 15%	15 - 30%	30 - 45%	≥ 45%
Age	1.2	1.0	0.6	0.5
Sex	1.2	1.0	0.6	0.5
Race	1.2	1.0	0.6	0.5
Hispanic origin (of any race)	1.2	1.0	0.6	0.5
Marital status	1.1	0.9	0.6	0.5
Household type and relationship	1.2	1.0	0.6	0.5
Children ever born	2.5	2.2	1.3	1.2
Work disability and mobility limitation status	1.2	1.0	0.6	0.5
Ancestry	1.8	1.5	1.0	0.8
Place of birth	1.9	1.6	1.0	0.9
Citizenship	1.7	1.4	1.0	0.8
Residence in 1985	1.9	1.7	1.0	0.9
Year of entry	1.3	1.0	0.6	0.5
Language spoken at home and ability to speak English	1.6	1.3	0.9	0.7
Educational attainment	1.3	1.1	0.6	0.5
School enrollment	1.6	1.4	1.0	0.8
Type of residence (urban/rural)	1.7	1.7	1.4	1.4
Household type	1.2	1.0	0.6	0.5
Family type	1.1	1.0	0.6	0.5
Group quarters	1.0	1.1	0.9	0.8
Subfamily type and presence of children	1.1	0.9	0.5	0.5
Employment status	1.2	1.0	0.6	0.5
Industry	1.2	1.0	0.6	0.5
Occupation	1.2	1.0	0.6	0.5
Class of worker	1.2	1.0	0.6	0.5
Hours per week and weeks worked in 1989	1.4	1.2	0.7	0.6
Number of workers in family	1.3	1.1	0.7	0.6
Place of work	1.4	1.2	0.8	0.6
Means of transportation to work	1.4	1.2	0.7	0.6
Travel time to work	1.3	1.1	0.6	0.5
Private vehicle occupancy	1.4	1.2	0.7	0.6
Time leaving to go to work	1.2	1.0	0.6	0.5
Type of income in 1989	1.3	1.1	0.6	0.5
Household income in 1989	1.1	1.0	0.6	0.5
Family income in 1989	1.1	1.0	0.6	0.5
Poverty status in 1989 (persons)	1.5	1.2	0.7	0.7
Poverty status in 1989 (families)	1.1	0.9	0.5	0.5
Armed forces and veteran status	1.4	1.1	0.7	0.6

Source: U.S. Bureau of the Census (1993). 1990 Census of Population: Social and Economic Characteristics: Vermont. Report Number 1990 CP-2-47. Page C-11.

A standard error based upon simple random sampling and a one-in-six sampling rate was computed thus:

$$S_{\mathrm{SRS}} = \sqrt{5\,\hat{Y}(1 - \hat{Y}/N)}$$

developed from standard formulas displayed in Cochran (1977).

For each data item within the weighting area, a design effect was computed as the ratio of the S_{RG} to S_{SRS}:

$$F = \frac{S_{\mathrm{RG}}}{S_{\mathrm{SRS}}}.$$

For a state report of sample data, the design effects for each data item were averaged across the weighting areas in the state. Then, a generalized design effect for each data item type (for example, all data items that dealt with occupation) was computed. The generalized design effect was weighted in favor of data items that had higher population estimates. Details on most of the procedures followed are available in a Census Bureau document (U.S. Bureau of the Census 1991). The same basic method was also used for sample data products in both the 1970 and 1980 census.

3. A HYPOTHETICAL EXAMPLE OF RANDOM GROUPS

Table 2 presents a hypothetical example of data that might have arisen from the random-groups method. For a weighting area in Vermont, weighted counts of whites and blacks are listed for the 25 random groups. In this hypothetical weighting area, there are no persons of other race. The standard errors assuming simple random sampling are the same for whites and blacks (as one would expect for a binomial variable). However, S_{RG} is much higher for the estimate of whites than the estimate of blacks. And, the design effect is nearly five times higher for the estimate of whites than the estimate of blacks. Since the generalized design effect computed for groups of data items was weighted in favor of data items that had higher population estimates, the generalized design effect computed for race for the state of Vermont was quite high.

Data on race were frequently included in 1990 U.S. census sample data products. Because race was asked of every census respondent (*i.e.*, it was a census 100-percent data item), and because the weighting process used by the Census Bureau effectively forced the sample estimates by race to match the 100-percent Census counts by race, the standard errors for estimates of race probably should have been considered to be zero. However, generalized design effects were still published by race, although set to arbitrary constants for all reports (rather than as computed by this method).

4. A MODIFIED APPROACH TO THE RANDOM GROUPS METHOD

A slight modification of the random groups method (essentially applying a ratio-estimation technique) can achieve much more satisfactory results in the estimation of standard errors. Rather than using $\hat{Y}_i$ as defined above for the estimate of the total for the i-th random group, one could instead use

$$\hat{L}_i = N X_i / W_i$$

Table 2
Hypothetical example of data that could have resulted from the Random Groups method used to estimate standard errors for census sample data.
For a weighting area in Vermont, people are asked their race.
A few (110) are black; most (2,518) are white.
A sampling rate of one-in-six is assumed ($N = 2{,}628$, $n = 438$).

Random Group	Weighted count of blacks*	Weighted count of whites*	Total weighted population count #
1	10	90	100
2	0	100	100
3	0	110	110
4	0	140	140
5	5	70	75
6	8	50	58
7	12	103	115
8	20	60	80
9	0	65	65
10	0	100	100
11	0	125	125
12	0	130	130
13	10	90	100
14	0	100	100
15	0	110	110
16	0	140	140
17	5	70	75
18	8	52	60
19	12	103	115
20	20	160	180
21	0	65	65
22	0	100	100
23	0	125	125
24	0	130	130
25	0	130	130
Sum of weighted counts ($\hat{Y}$)	110	2,518	2,628
S_{RG}	145.98	687.96	
S_{SRS}	22.96	22.96	
F	6.36	29.96	

* The first 25 figures in this column represent X_i for the i-th random group under the modified random groups method. Multiplying the figure by 25 yields $\hat{Y}_i$ for the random groups method employed by the U.S. Bureau of the Census.

\# The first 25 figures in this column represent W_i under the modified random groups method.

where X_i represents the weighted count for the data item within the i-th random group, W_i is the weighted count of all persons in the i-th random group, and N represents the census count of persons in the weighting area. The modified random groups standard error estimate is then

$$S_L = \sqrt{(1 - n/N) \sum_{i=1}^{25} \frac{(\hat{L}_i - \hat{Y})^2}{24}}.$$

Using this method, S_L is 160.78 for both blacks and whites in the hypothetical weighting area of Table 1 (close to the value of S_{RG} for blacks). In this case, the requirement for standard error estimates for both responses for a binomial variable to be identical is met. Moreover, if all sample units have the same response for some variable, S_L becomes zero, whereas S_{RG} only becomes zero when each random group has the same weighted count.

This modified standard error estimation procedure could be useful for researchers who do not have access to any of the many computer programs now available for computing estimates from sample data (such as SUDAAN, STATA, PC-CARP, VPLX, *etc.*). In addition, the U.S. Bureau of the Census ought to consider modifying its approach for estimating standard errors for sample data from the 2000 census. Moreover, with the U.S. Bureau of the Census' current emphasis on quality management, the U.S. Bureau of the Census may wish to poll users of sample data products to determine how useful the presentation of standard errors (through design effects) was to them, and involve a number of the data users in improving the presentation of standard errors for the next census.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques* (third edition). New York: John Wiley & Sons.

WOLTER, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

U.S. BUREAU OF THE CENSUS (1991). Computer Specifications for the 1990 Decennial Census Variance Estimation Operation. STSD Decennial Census Memorandum Series #Z-65.

U.S. BUREAU OF THE CENSUS (1993). Appendix C. Accuracy of the Data. Pp. C-1 to C-11 in 1990 Census of Population: Social and Economic Characteristics: Vermont. Bureau of the Census Document 1990 CP-2-47.

Survey Methodology, June 1997
Vol. 23, No. 1, pp. 77-78
Statistics Canada

A Simple Derivation of the Linearization of the Regression Estimator

KEES ZEELENBERG[1]

ABSTRACT

We show how the use of matrix calculus can simplify the derivation of the linearization of the regression coefficient estimator and the regression estimator.

KEY WORDS: Matrix calculus; Regression estimator; Taylor expansion.

1. INTRODUCTION

Design-based sampling variances of non-linear statistics are often calculated by means of a linear approximation obtained by a Taylor expansion; examples are the variances of the general regression coefficient estimator and the regression estimator. The linearizations usually need some complicated differentiations. The purpose of this paper is to show how matrix calculus can simplify these derivations, to the extent that even the Taylor expansion of the regression coefficient estimator can be derived in one line, which should be compared with the nearly one page that Särndal *et al.* (1992, p. 205-206) need. To be honest, the use of matrix calculus requires some more machinery to be set up, which is not needed for traditional methods. However this set-up can be regarded as an investment; once it has been learned, it can be used fruitfully in many other applications. After this paper had been written, Binder (1996) appeared, in which similar techniques are used to derive variances by means of linearization. The present paper can be seen as a pedagogical note, in which the use of differentials is exposed.

2. MATRIX DIFFERENTIALS

2.1 Introduction

We will use the matrix calculus by means of differentials, as set out by Magnus and Neudecker (1988); this calculus differs somewhat from the usual methods, which focus on derivatives instead of differentials. Therefore in this section we will briefly describe the definitions and properties of differentials (see Zeelenberg 1993, for a more extensive survey). We first define differentials for vector functions, and then generalize to matrix functions.

2.2 Vector Functions

Let f be a function from an open set $S \subset \mathbb{R}^m$ to $\mathbb{R}^n$; let x_0 be a point in S. The function f is *differentiable* at x_0 if there exists a real $n \times m$-matrix A, depending on x_0, such that for any $u \in \mathbb{R}^m$ for which $x_0 + u \in S$, there holds

$$f(x_0 + u) = f(x_0) + A_{x_0} u + o(u), \tag{1}$$

where $o(u)$ is a function such that $\lim_{|u| \to 0} |o(u)|/|u| = 0$; the matrix A is called the *first derivative* of f at x_0; it is denoted as $Df(x_0)$ or $\partial f/\partial(x')\rfloor_{x=x_0}$. The derivative Df is equal to the matrix of partial derivatives, *i.e.*, $Df(x)_{ij} = \partial f_i/\partial x_j$. The linear function $df_{x_0}: \mathbb{R}^m \mapsto \mathbb{R}^n$ defined by $df_{x_0}: u \mapsto A_{x_0} u$ is called the *differential* of f at x_0. Usually we write dx instead of u so that $df_{x_0}(dx) = A_{x_0} dx$. From (1) we see that the differential corresponds to the linear part of the function, which can also be written as

$$y - y_0 = A_{x_0}(x - x_0),$$

where $y_0 = f(x_0)$. Therefore the differential of a function is the linearization of the function: it is the equation of the hyperplane through the origin that is parallel to the hyperplane tangent to the graph of f at x_0; so the linearized function can be written as

$$f(x) \doteq f(x_0) + A_{x_0}(x - x_0). \tag{2}$$

Alternatively, if B is a matrix such that $df_{x_0}(dx) = Bdx$, then B is the derivative of f at x_0 and contains the partial derivatives of f at x_0. This one-to-one relationship between differentials and derivatives is very useful, since differentials are easy to manipulate.

Finally, we usually omit the subscript 0 in x_0, so that we write $df = A_x dx$.

2.3 Matrix Functions

A matrix function F from an open set $S \subset \mathbb{R}^{m \times n}$ to $\mathbb{R}^{p \times q}$ is differentiable if vec F is differentiable. The derivative DF is the derivative of vec F with respect to vec X, and is also denoted by $\partial \operatorname{vec} F/\partial(\operatorname{vec} X)'$. The differential dF is the matrix function defined by vec $dF_{X_0}(U) = A_{X_0} \operatorname{vec} U$.

[1] Kees Zeelenberg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands.

2.4 Properties of Differentials

Let A be a matrix of constants, F and G differentiable matrix functions, and α a real scalar. Then the following properties are easily proved:

$$dA = 0, \tag{3}$$

$$d(\alpha F) = \alpha dF, \tag{4}$$

$$d(F + G) = dF + dG, \tag{5}$$

$$d(FG) = (dF)G + F(dG), \tag{6}$$

$$dF^{-1} = -F^{-1}(dF)F^{-1}. \tag{7}$$

The last property can be proved by taking the differential of $FF^{-1} = I$ and rearranging.

3. LINEARIZATION OF THE REGRESSION COEFFICIENT ESTIMATOR

The π-estimator (Horvitz-Thompson estimator) of the finite population regression coefficient (*cf.* Särndal *et al.* 1992, section 5.10) is

$$\hat{B} = \hat{T}^{-1}\hat{t}, \tag{8}$$

where

$$\hat{T} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k},$$

$$\hat{t} = \sum_{k \in s} \frac{x_k y_k}{\pi_k},$$

y_k is the variable of interest for individual k, x_k is the vector with the auxiliary variables for individual k, π_k is the inclusion probability for individual k, and s denotes the sample.

Taking the total differential of (8), using properties (6) and (7), and evaluating at the point where $\hat{T} = T, \hat{t} = t$, we get

$$d\hat{B} = -T^{-1}(d\hat{T})T^{-1}t + T^{-1}(d\hat{t}). \tag{9}$$

Because of the connection between differentials and linear approximation, as given in equation (2), it immediately follows that (9) corresponds to the linearization of the regression coefficient estimator:

$$\hat{B} \doteq B - T^{-1}(\hat{T} - T)T^{-1}t + T^{-1}(\hat{t} - t) = B + T^{-1}(\hat{t} - \hat{T}B),$$

where $B = T^{-1}t$.

4. LINEARIZATION OF THE REGRESSION ESTIMATOR

The regression estimator of a population total is (*cf.* Särndal *et al.* 1992, section 6.6)

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}, \tag{10}$$

where $\hat{t}_{y\pi}$ is the π-estimator of the variable of interest, t_x is the vector with the population totals of the auxiliary variables, $\hat{t}_{x\pi}$ is the vector with the π-estimators of the auxiliary variables, and $\hat{B}$ is the estimator of the regression coefficient of the auxiliary variables on the variable of interest. Taking the total differential of (10), using properties (3) and (6), and evaluating at the point where $\hat{t}_{y\pi} = t_y, \hat{t}_{x\pi} = t_x$, and $\hat{B} = B$, we get the linear approximation of the regression estimator

$$d\hat{t}_{yr} = d\hat{t}_{y\pi} - (d\hat{t}_{x\pi})'B,$$

so that

$$\hat{t}_{yr} \doteq t_y + \hat{t}_{y\pi} - t_y + (t_x - \hat{t}_{x\pi})'B = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'B.$$

Note that for the linearization of the regression estimator we do not need that of the regression coefficient estimator B.

ACKNOWLEDGEMENTS

I wish to thank Jeroen Pannekoek, Jos de Ree, Robbert Renssen, two referees, and an Associate Editor for their comments. The views expressed in this article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

REFERENCES

BINDER, D.A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22, 17-22.

MAGNUS, J.R., and NEUDECKER, H. (1988). *Matrix Differential Calculus*. New York: Wiley.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

ZEELENBERG, C. (1993). *A Survey of Matrix Differentiation*. Research Paper, Department of Statistical Methods, Statistics Netherlands, Voorburg.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 1, March/mars 1997

J.N.K. RAO
Developments in sample survey theory: an appraisal

T.M. Fred SMITH
Social surveys and social science

Feifang HU
The asymptotic properties of the maximum relevance weighted likelihood estimators

R.R. SITTER and J.N.K. RAO
Imputation for missing values and corresponding variance estimation

Patrick J. FARRELL, Brenda MacGIBBON and Thomas J. TOMBERLIN
Bootstrap adjustments for empirical Bayes interval estimates of small area proportions

D.A.S. FRASER, N. REID and A. WONG
Simple and accurate inference for the mean of the gamma model

Jianguo SUN and David E. MATTHEWS
A random-effect regression model for medical follow-up studies

Philippe CAPÉRAÀ and Ana Isabel Garralda GUILLEM
Taux de résistance des tests de rang d'indépendance

Volume 25, No. 2, June/juin 1997

X. Joan HU and Jerald F. LAWLESS
Pseudolikelihood estimation in a class of problems with response-related missing covariates

Irwin GUTTMAN and George D. PAPANDONATOS
A Bayesian approach to a reliability problem: theory, analysis and interesting numerics

R.J. OHARA HINES
Fitting generalized linear models to retrospectively sampled clusters with categorical responses

R.R. SITTER and I. FAINARU
Optimal designs for the logit and probit models for binary data

Boxin TANG and C.F.J. WU
A method for constructing supersaturated designs and its $E(s^2)$ optimality

Shu YAMADA and Dennis K.J. LIN
Supersaturated design including an orthogonal base

A.G. BENN and R.J. KULPBERGER
Integrated marked Poisson processes with application to image correlation spectroscopy

Khalid El HIMDI and Roch ROY
Tests for the non-correlation of two multivariate ARMA time series

John J. SPINELLI and Michael A. STEPHENS
Cramér-von Mises tests of fit for the Poisson distribution

Thomas W. O'GORMAN
An adaptive test for the one-way layout

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 12, Number 4, 1996

Derivation and Properties of the X11ARIMA and Census X11 Linear Filters
Estela Bee Dagum, Norma Chhab, and Kim Chiu 329
Correcting Unit Nonresponse via Response Modeling and Raking in the California Tobacco Survey
Charles C. Berry, Shirley W. Cavin, and John P. Pierce 349
Multiple Workloads per Stratum Designs
Lynn Weidmann and Lawrence R. Ernst 365
Neural Network Imputation Applied to the Norwegian 1990 Population Census Data
Svein Nordbotten 385
Modeling Income in the U.S. Consumer Expenditure Survey
Geoffrey D. Paulin and Elizabeth M. Sweet 403
The Survey Reinterview: Respondent Perceptions and Response Strategies
Johnny Blair and Seymour Sudman 421
Corrigendum 427
Book Reviews 429
Editorial Collaborators 441
Index to Volume 12, 1996 445

Volume 13, Number 1, 1997

Who Lives Here? Survey Undercoverage and Household Roster Questions
Roger Tourangeau, Gary Shapiro, Anne Kearney, and Lawrence Ernst 1
Suggestive Interviewer Behaviour in Surveys: An Experimental Study
Johannes H. Smit, Wil Dijkstra, and Johannes van der Zouwen 19
Effects of Post-Stratification on the Estimates of the Finnish Labour Force Survey
Kari Djerf 29
Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach
Milorad S. Kovačević and David A. Binder 41
Issues in the Use of a Plant-Capture Method for Estimating the Size of the Street Dwelling Population
Elizabeth Martin, Eugene Laska, Kim Hopper, Morris Meisner, and Joe Wanderling 59
A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data
Stephen E. Fienberg, Udi E. Makov, and Ashish P. Sanil 75
Book Review 91
In Other Journals 101

Volume 13, Number 2, 1997

Evaluation of a Reconstruction of the Adjusted 1990 Census for Florida
Michael M. Meyer and Joseph B. Kadane 103
Individual Diaries and Expense Documents in the Italian Consumer Expenditure Survey
Carlo Filippucci and Maria Rosaria Ferrante 113
Testing of Distribution Functions from Complex Sample Surveys
Abba M. Krieger and Danny Pfeffermann 123
Estimating Consumer Price Indices for Small Reference Populations
Martin Boon and Jan de Haan 143
Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers
Norbert Schwarz and Tracy Wellens 159
Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality
Bärbel Knäuper, Robert F. Belli, Daniel H. Hill, and A. Regula Herzog 181

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 12, Number 4, 1996

Derivation and Properties of the X11ARIMA and Census X11 Linear Filters
Estela Bee Dagum, Norma Chhab, and Kim Chiu 329
Correcting Unit Nonresponse via Response Modeling and Raking in the California Tobacco Survey
Charles C. Berry, Shirley W. Cavin, and John P. Pierce 349
Multiple Workloads per Stratum Designs
Lynn Weidmann and Lawrence R. Ernst 365
Neural Network Imputation Applied to the Norwegian 1990 Population Census Data
Svein Nordbotten 385
Modeling Income in the U.S. Consumer Expenditure Survey
Geoffrey D. Paulin and Elizabeth M. Sweet 403
The Survey Reinterview: Respondent Perceptions and Response Strategies
Johnny Blair and Seymour Sudman 421
Corrigendum 427
Book Reviews 429
Editorial Collaborators 441
Index to Volume 12, 1996 445

Volume 13, Number 1, 1997

Who Lives Here? Survey Undercoverage and Household Roster Questions
Roger Tourangeau, Gary Shapiro, Anne Kearney, and Lawrence Ernst 1
Suggestive Interviewer Behaviour in Surveys: An Experimental Study
Johannes H. Smit, Wil Dijkstra, and Johannes van der Zouwen 19
Effects of Post-Stratification on the Estimates of the Finnish Labour Force Survey
Kari Djerf 29
Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach
Milorad S. Kovačević and David A. Binder 41
Issues in the Use of a Plant-Capture Method for Estimating the Size of the Street Dwelling Population
Elizabeth Martin, Eugene Laska, Kim Hopper, Morris Meisner, and Joe Wanderling 59
A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data
Stephen E. Fienberg, Udi E. Makov, and Ashish P. Sanil 75
Book Review 91
In Other Journals 101

Volume 13, Number 2, 1997

Evaluation of a Reconstruction of the Adjusted 1990 Census for Florida
Michael M. Meyer and Joseph B. Kadane 103
Individual Diaries and Expense Documents in the Italian Consumer Expenditure Survey
Carlo Filippucci and Maria Rosaria Ferrante 113
Testing of Distribution Functions from Complex Sample Surveys
Abba M. Krieger and Danny Pfeffermann 123
Estimating Consumer Price Indices for Small Reference Populations
Martin Boon and Jan de Haan 143
Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers
Norbert Schwarz and Tracy Wellens 159
Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality
Bärbel Knäuper, Robert F. Belli, Daniel H. Hill, and A. Regula Herzog 181

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 1, March/mars 1997

J.N.K. RAO
Developments in sample survey theory: an appraisal

T.M. Fred SMITH
Social surveys and social science

Feifang HU
The asymptotic properties of the maximum relevance weighted likelihood estimators

R.R. SITTER and J.N.K. RAO
Imputation for missing values and corresponding variance estimation

Patrick J. FARRELL, Brenda MacGIBBON and Thomas J. TOMBERLIN
Bootstrap adjustments for empirical Bayes interval estimates of small area proportions

D.A.S. FRASER, N. REID and A. WONG
Simple and accurate inference for the mean of the gamma model

Jianguo SUN and David E. MATTHEWS
A random-effect regression model for medical follow-up studies

Philippe CAPÉRAÀ and Ana Isabel Garralda GUILLEM
Taux de résistance des tests de rang d'indépendance

Volume 25, No. 2, June/juin 1997

X. Joan HU and Jerald F. LAWLESS
Pseudolikelihood estimation in a class of problems with response-related missing covariates

Irwin GUTTMAN and George D. PAPANDONATOS
A Bayesian approach to a reliability problem: theory, analysis and interesting numerics

R.J. OHARA HINES
Fitting generalized linear models to retrospectively sampled clusters with categorical responses

R.R. SITTER and I. FAINARU
Optimal designs for the logit and probit models for binary data

Boxin TANG and C.F.J. WU
A method for constructing supersaturated designs and its $E(s^2)$ optimality

Shu YAMADA and Dennis K.J. LIN
Supersaturated design including an orthogonal base

A.G. BENN and R.J. KULPBERGER
Integrated marked Poisson processes with application to image correlation spectroscopy

Khalid El HIMDI and Roch ROY
Tests for the non-correlation of two multivariate ARMA time series

John J. SPINELLI and Michael A. STEPHENS
Cramér-von Mises tests of fit for the Poisson distribution

Thomas W. O'GORMAN
An adaptive test for the one-way layout

2.4 Propriétés des différentielles

Soit A, une matrice de constantes; F et G, des fonctions matricielles dérivables, et α, une valeur scalaire réelle. De ce qui précède, on peut facilement prouver les propriétés suivantes:

$$dA = 0, \tag{3}$$

$$d(\alpha F) = \alpha dF, \tag{4}$$

$$d(F + G) = dF + dG, \tag{5}$$

$$d(FG) = (dF)G + F(dG), \tag{6}$$

$$dF^{-1} = -F^{-1}(dF)F^{-1}. \tag{7}$$

Pour prouver la dernière propriété, il suffit de prendre la différentielle de $FF^{-1} = I$ et de la restructurer.

3. LINÉARISATION DE L'ESTIMATEUR DU COEFFICIENT DE RÉGRESSION

L'estimateur π (estimateur de Horvitz-Thompson) du coefficient de régression d'une population finie (lire Särndal et coll. 1992, partie 5.10) est

$$\hat{B} = \hat{T}^{-1}\hat{t}, \tag{8}$$

où

$$\hat{T} = \sum_{k \in s} \frac{x_k x_k'}{\pi_k},$$

$$\hat{t} = \sum_{k \in s} \frac{x_k y_k}{\pi_k},$$

y_k est la variable à laquelle on s'intéresse pour chaque k, x_k est le vecteur de chaque k avec les variables auxiliaires, π_k est la probabilité d'inclusion de chaque k et s représente l'échantillon.

Lorsqu'on prend la différentielle totale de (8) au moyen des propriétés (6) et (7) et l'évalue au point où $\hat{T} = T, \hat{t} = t$, on obtient

$$d\hat{B} = -T^{-1}(d\hat{T})T^{-1}t + T^{-1}(d\hat{t}). \tag{9}$$

Étant donné le lien entre les différentielles et l'approximation linéaire signalé à l'équation (2), il s'ensuit que (9) correspond à la linéarisation de l'estimateur du coefficient de régression:

$$\hat{B} \doteq B - T^{-1}(\hat{T}-T)T^{-1}t + T^{-1}(\hat{t}-t) = B + T^{-1}(\hat{t}-\hat{T}B),$$

où $B = T^{-1}t$.

4. LINÉARISATION DE L'ESTIMATEUR DE RÉGRESSION

L'estimateur de régression d'un chiffre de population est (lire Särndal et coll. 1992, partie 6.6)

$$\hat{t}_{yr} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'\hat{B}, \tag{10}$$

où $\hat{t}_{y\pi}$ est l'estimateur π de la variable à laquelle on s'intéresse, t_x est le vecteur du chiffre de population pour les variables auxiliaires, $\hat{t}_{x\pi}$ est le vecteur des estimateurs π des variables auxiliaires et $\hat{B}$ est l'estimateur du coefficient de régression des variables auxiliaires de la variable à laquelle on s'intéresse. Lorsqu'on calcule le différentielle totale de (10) grâce aux propriétés (3) et (6) et l'évalue au point où $\hat{t}_{y\pi} = t_y, \hat{t}_{x\pi} = t_x$, et $\hat{B} = B$, on obtient l'approximation linéaire de l'estimateur de régression

$$d\hat{t}_{yr} = d\hat{t}_{y\pi} - (d\hat{t}_{x\pi})'B,$$

si bien que

$$\hat{t}_{yr} \doteq t_y + \hat{t}_{y\pi} - t_y + (t_x - \hat{t}_{x\pi})'B = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})'B.$$

Notons qu'il n'est pas nécessaire de linéariser l'estimateur du coefficient de régression B pour linéariser l'estimateur de régression.

REMERCIEMENTS

L'auteur tient à remercier Jeroen Pannekoek, Jos de Ree, Robbert Renssen, deux arbitres et un rédacteur associé pour leurs précieux commentaires. Le point de vue exprimé dans cet article n'engage que l'auteur et ne reflète pas nécessairement les politiques du Statistics Netherland.

BIBLIOGRAPHIE

BINDER, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette». *Techniques d'enquête*, 22, 17-22.

MAGNUS, J.R., et NEUDECKER, H. (1988). *Matrix Differential Calculus*. New York: Wiley.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

ZEELENBERG, C. (1993). A Survey of Matrix Differentiation. Research Paper, Department of Statistical Methods. Voorburg: Statistics Netherlands.

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 85-86
Statistique Canada

Dérivation simple de l'estimateur de régression par linéarisation

KEES ZEELENBERG[1]

RÉSUMÉ

L'auteur explique comment recourir au calcul matriciel pour simplifier la dérivation de l'estimateur du coefficient de régression et de l'estimateur de régression par linéarisation.

MOTS CLÉS: Calcul matriciel; estimateur de régression; développement de Taylor.

1. INTRODUCTION

On calcule souvent la variance des statistiques non linéaires obtenues par échantillonnage au moyen d'une méthode d'approximation linéaire faisant appel à l'expansion de Taylor. La variance de l'estimateur général du coefficient de régression et de l'estimateur de régression en est un exemple. Habituellement, la linéarisation exige des dérivations complexes. Le présent article a pour but de montrer comment le calcul matriciel peut simplifier les dérivations de ce genre, au point où l'on pourrait dériver l'expansion de Taylor de l'estimateur du coefficient de régression en une ligne plutôt qu'en près d'une page ainsi que le proposent Särndal et ses collaborateurs (1992, p. 205-206). Il convient néanmoins de souligner que le calcul matriciel exige certains préparatifs, inutiles avec les méthodes classiques. Quoi qu'il en soit, ces préparatifs peuvent être considérés comme un placement puisqu'une fois connus, on peut en faire un usage fructueux dans de nombreuses applications. Après la rédaction de cet article, Binder (1996) en a publié un autre présentant des techniques analogues pour dériver les variances par linéarisation. Le présent article peut être considéré comme une note pédagogique expliquant comment utiliser les différentielles.

2. DIFFÉRENTIELLES MATRICIELLES

2.1 Introduction

Nous recourrons au calcul matriciel en faisant appel aux différentielles selon la description de Magnus et Neudecker (1988). Cette méthode de calcul s'écarte quelque peu des méthodes habituelles, qui reposent plus sur les dérivées que les différentielles. C'est pourquoi nous commencerons par examiner brièvement la définition et les propriétés des différentielles (lire Zeelenberg 1993, pour plus d'explications). Nous définirons d'abord les différentielles à l'égard des fonctions vectorielles avant de généraliser la définition aux fonctions matricielles.

2.2 Fonctions vectorielles

Soit f, une fonction de l'ensemble ouvert $S \subset \mathbb{R}^m$ à $\mathbb{R}^n$ et x_0, un point de S. La fonction f est *dérivable* à x_0 s'il existe une vraie matrice $n \times m$ A, dépendant de x_0, de sorte que pour chaque valeur $u \in \mathbb{R}^m$ pour laquelle $x_0 + u \in S$,

$$f(x_0 + u) = f(x_0) + A_{x_0} u + o(u), \tag{1}$$

où $o(u)$ est une fonction selon laquelle $\lim_{|u| \to 0} |o(u)|/|u| = 0$. La matrice A s'appelle *dérivée première* de f à x_0 et est notée $Df(x_0)$ ou $\partial f/\partial(x')\rfloor_{x=x_0}$. La dérivée Df est égale à la matrice des dérivées partielles, c'est-à-dire $Df(x)_{ij} = \partial f_i/\partial x_j$. La fonction linéaire $df_{x_0}: \mathbb{R}^m \mapsto \mathbb{R}^n$ définie par $df_{x_0}: u \mapsto A_{x_0} u$ est la *dérivée* de f à x_0. On écrit dx plus souvent que u, de sorte que $df_{x_0}(dx) = A_{x_0} dx$. D'après (1), on se rend compte que la différentielle correspond à la partie linéaire de la fonction, qu'on peut aussi exprimer par

$$y - y_0 = A_{x_0}(x - x_0),$$

ou $y_0 = f(x_0)$. La différentielle d'une fonction est donc la linéarisation de cette fonction, c'est-à-dire l'équation de l'hyperplan qui coupe l'origine parallèlement à l'hyperplan tangent à la courbe de f à x_0. La fonction linéarisée est donc

$$f(x) \doteq f(x_0) + A_{x_0}(x - x_0). \tag{2}$$

Par ailleurs, si B est une matrice telle que $df_{x_0}(dx) = Bdx$, alors B est la dérivée de f à x_0 et comprend les dérivées partielles de f à x_0. Cette relation biunivoque entre différentielles et dérivées s'avère fort utile, car les différentielles sont faciles à manipuler.

Enfin, on omet habituellement l'indice 0 dans x_0, ce qui donne $df = A_x dx$.

2.3 Fonctions matricielles

Une fonction matricielle F d'un ensemble ouvert $S \subset \mathbb{R}^{m \times n}$ à $\mathbb{R}^{p \times q}$ est dérivable si vec F est dérivable. La dérivée DF correspond à la dérivée de vec F par rapport à vec X et est notée ∂ vec $F/\partial(\text{vec } X)'$. La différentielle dF représente la fonction matricielle définie par vec $dF_{X_0}(U) = A_{X_0} \text{vec } U$.

---

[1] Kees Zeelenberg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands.

Tableau 2
Exemple hypothétique de données qu'on aurait pu obtenir au moyen de la méthode des groupes aléatoires pour estimer l'erreur-type des données échantillonnées au recensement. Pour une région de pondération du Vermont, on a demandé aux répondants de préciser leur race. Quelques personnes (110) sont noires; la plupart (2,518) sont blanches. On suppose un taux d'échantillonnage d'une personne sur six ($N = 2{,}628$, $n = 438$).

Groupe aléatoire	Population pondéré de Noirs*	Population pondéré de Blancs*	Total du chiffre de la population pondérée #
1	10	90	100
2	0	100	100
3	0	110	110
4	0	140	140
5	5	70	75
6	8	50	58
7	12	103	115
8	20	60	80
9	0	65	65
10	0	100	100
11	0	125	125
12	0	130	130
13	10	90	100
14	0	100	100
15	0	110	110
16	0	140	140
17	5	70	75
18	8	52	60
19	12	103	115
20	20	160	180
21	0	65	65
22	0	100	100
23	0	125	125
24	0	130	130
25	0	130	130
Somme des chiffres pondérés ($\hat{Y}$)	110	2,518	2,628
S_{RG}	145.98	687.96	
S_{SRS}	22.96	22.96	
F	6.36	29.96	

* Les 25 permiers chiffres de la colonne correspondent à la valeur X_i du i-ième groupe aléatoire, selon la méthode modifiée des groupes aléatoires. En multipliant ce chiffre par 25, on obtient $\hat{Y}_i$ selon la méthode des groupes aléatoires utilisée par le U.S. Bureau of the Census.

\# Les 25 premiers chiffres de cette colonne représentent W_i, selon la méthode modifiée des groupes aléatoires.

4. NOUVELLE APPROCHE À LA MÉTHODE DES GROUPES ALÉATOIRES

On obtiendrait des résultats beaucoup plus satisfaisants dans l'estimation de l'erreur-type en modifiant légèrement la méthode des groupes aléatoires (essentiellement en recourant à une technique d'estimation par ratio). Au lieu d'utiliser $\hat{Y}_i$ tel que défini ci-dessus pour estimer le total du i-ième groupe aléatoire, on pourrait recourir à l'équation:

$$\hat{L}_i = NX_i/W_i$$

où X_i représente la population pondérée pour l'élément de donnée du i-ième groupe aléatoire, W_i correspond au chiffre pondéré de population du i-ième groupe aléatoire et N indique la population recensée dans la région de pondération. L'erreur-type estimative obtenue avec la méthode modifiée des groupes aléatoires serait donc

$$S_L = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{L}_i - \hat{Y})^2}{24}}.$$

Grâce à cette méthode, la valeur de S_L pour les Noirs et les Blancs dans la région de pondération hypothétique du tableau 1 est égale à 160.78 (ce qui s'approche de la valeur de S_{RG} pour les Noirs). Bref, l'erreur-type est la même pour les deux réponses à la question binomiale. D'autre part, si toutes les unités échantillonnées fournissent la même réponse, S_L est égal à zéro, alors que S_{RG} n'obtient la valeur nulle que lorsque chaque groupe aléatoire a la même population pondérée.

Cette nouvelle méthode d'estimation de l'erreur-type pourrait s'avérer utile aux chercheurs qui n'ont pas accès à l'un des nombreux logiciels dont on peut désormais se servir pour effectuer des estimations à partir des données d'échantillon (par exemple SUDAAN, STATA, PC-CARP, VPLX, *etc.*). Par ailleurs, le U.S. Bureau of the Census devrait envisager de modifier sa méthode d'estimation de l'erreur-type pour les données d'échantillon en prévision du recensement de l'an 2000. Enfin, compte tenu de l'intérêt qu'on porte présentement à la gestion de la qualité au Census Bureau, ce dernier pourrait vouloir sonder les utilisateurs de produits de données afin d'établir l'utilité de la présentation actuelle des erreurs-types (par le truchement des effets du plan d'échantillonnage) et demander à un certain nombre d'entre eux de participer à un exercice ayant pour but d'améliorer la manière dont les erreurs-types seront présentées au prochain recensement.

BIBLIOGRAPHIE

COCHRAN, W. G. (1977). *Sampling Techniques* (3ième édition). John Wiley & Sons: New York.

WOLTER, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag: New York.

U.S. BUREAU OF THE CENSUS (1991). Computer Specifications for the 1990 Decennial Census Variance Estimation Operation. STSD Decennial Census Memorandum Series #Z-65.

U.S. BUREAU OF THE CENSUS (1993). Appendix C. Accuracy of the Data. Pp. C-1 à C-11 dans 1990 Census of Population: Social and Economic Characteristics: Vermont. Bureau of the Census Document 1990 CP-2-47.

Tableau 1
Effects du plan d'échantillonnage du recensement américain de 1990 indiqués pour les données échantillonnées au Vermont

Caratéristique	Pourcentage de personnes ou d'unités de logement dans l'échantillon			
	< 15%	15 - 30%	30 - 45%	≥ 45%
Âge	1.2	1.0	0.6	0.5
Sexe	1.2	1.0	0.6	0.5
Race	1.2	1.0	0.6	0.5
Ascendance hispanique (peu importe la race)	1.2	1.0	0.6	0.5
État civil	1.1	0.9	0.6	0.5
Type de ménage et relation	1.2	1.0	0.6	0.5
Enfants	2.5	2.2	1.3	1.2
Limitation de mobilité et incapacité liée au travail	1.2	1.0	0.6	0.5
Origines	1.8	1.5	1.0	0.8
Lieu de naissance	1.9	1.6	1.0	0.9
Citoyenneté	1.7	1.4	1.0	0.8
Lieu de rédience en 1985	1.9	1.7	1.0	0.9
Année d'entrée	1.3	1.0	0.6	0.5
Langue parlée à domicile et aptitude à parler l'anglais	1.6	1.3	0.9	0.7
Scolarité	1.3	1.1	0.6	0.5
Inscription à l'école	1.6	1.4	1.0	0.8
Type de résidence (urbaine/rurale)	1.7	1.7	1.4	1.4
Type de ménage	1.2	1.0	0.6	0.5
Type de famille	1.1	1.0	0.6	0.5
Logement collectif	1.0	1.1	0.9	0.8
Type de sous-famille et présence d'enfants	1.1	0.9	0.5	0.5
Situation d'emploi	1.2	1.0	0.6	0.5
Secteur d'activité	1.2	1.0	0.6	0.5
Profession	1.2	1.0	0.6	0.5
Type de travailleur	1.2	1.0	0.6	0.5
Heures de travail par semaine et semaines de travail en 1989	1.4	1.2	0.7	0.6
Nombre de travailleurs	1.3	1.1	0.7	0.6
Lieu de travail	1.4	1.2	0.8	0.6
Moyen de transport au lieu de travail	1.4	1.2	0.7	0.6
Durée du déplacement	1.3	1.1	0.6	0.5
Usage d'un véhicule privé	1.4	1.2	0.7	0.6
Heure de départ pour le travail	1.2	1.0	0.6	0.5
Type de revenu en 1989	1.3	1.1	0.6	0.5
Revenu du ménage en 1989	1.1	1.0	0.6	0.5
Revenu de la famille en 1989	1.1	1.0	0.6	0.5
Indice de pauvreté en 1989	1.5	1.2	0.7	0.7
Indice de pauvreté en 1989	1.1	0.9	0.5	0.5
Forces armées et status d'ancien combattant	1.4	1.1	0.7	0.6

Source: U.S. Bureau of the Census (1993). 1990 Census of Population: Social and Economic Characteristics: Vermont. Rapport numéro 1990 CP-2-47. Page C-11.

où n correspond au nombre non pondéré de personnes de l'échantillon dans la région de pondération, N représente la population recensée dans la région de pondération, $\hat{Y}_i$ indique le total estimatif de l'élément de donnée obtenu en multipliant la population pondérée de l'élément en question du i-ième groupe aléatoire par 25 et $\hat{Y}$ représente le chiffre de population pondéré de l'élément de donnée (à savoir l'estimation de l'échantillon) pour la région de pondération.

L'erreur-type a été calculée selon la méthode d'échantillonnage aléatoire simple et un taux d'échantillonnage d'une personne sur six comme suit:

$$S_{SRS} = \sqrt{5\ \hat{Y}\ (1 - \hat{Y}/N)}.$$

Cette équation dérive des formules habituelles qu'on trouve dans Cochran (1977).

On a mesuré l'effet du plan d'échantillonnage sur chaque élément de donnée dans la région de pondération sous la forme du ratio entre S_{RG} et S_{SRS}:

$$F = \frac{S_{RG}}{S_{SRS}}.$$

Afin de produire un rapport sur les données échantillonnées pour l'État, on a établi les effets moyens du plan d'échantillonnage sur chaque élément de donnée pour l'ensemble des régions de pondération de l'État. On a ensuite calculé l'effet généralisé du plan d'échantillonnage sur chaque élément de donnée (par exemple, tous les éléments de donnée se rapportant à la profession). L'effet généralisé du plan d'échantillonnage a été pondéré en faveur des éléments de donnée pour lesquels la population estimative était plus élevée. Un document du Census Bureau (U.S. Bureau of the Census 1991) donne des précisions sur la plupart des méthodes suivies. On a recouru à la même approche fondamentale pour les produits de données des recensements de 1970 et de 1980.

3. EXEMPLE HYPOTHÉTIQUE DE GROUPES ALÉATOIRES

Le tableau 2 donne un exemple hypothétique de données qu'on aurait pu obtenir grâce à la méthode des groupes aléatoires. La population pondérée de Blancs et de Noirs est indiquée pour les 25 groupes aléatoires d'une région de pondération du Vermont. Dans cette région de pondération fictive, on ne compte aucune personne d'une autre race.

En supposant un échantillonnage simple aléatoire, l'erreur-type est la même pour les deux groupes (ainsi qu'on pourrait s'y attendre avec une variable binomiale). Toutefois, la valeur de S_{RG} est beaucoup plus élevée pour la population blanche que pour la population noire. En outre, l'effet du plan d'échantillonnage est près de cinq fois plus élevé pour les Blancs que pour les Noirs. Puisque l'effet généralisé du plan d'échantillonnage sur les groupes d'éléments de donnée a été pondéré en faveur des éléments de donnée pour lesquels la population estimative était plus élevée, l'effet généralisé du plan d'échantillonnage obtenu pour la race était très important au Vermont.

Les produits de données sur le recensement de 1990 des États-Unis présentaient fréquemment les données sur la race. Puisqu'on a demandé à chaque répondant d'indiquer sa race (il s'agissait d'un élément de donnée couvrant la totalité des personnes recensées) et puisque la méthode de pondération dont le Census Bureau s'est servi a contraint les estimations selon la race à correspondre au chiffre de population global du recensement selon la race, on aurait sans doute dû considéré que les erreurs-types de l'estimation de la race étaient égales à zéro. Toutefois, on a continué de diffuser les effets généralisés du plan d'échantillonnage selon la race, même si des constantes ont été arbitrairement établies dans tous les rapports (on ne les a pas calculés au moyen de la méthode indiquée).

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 81-83
Statistique Canada

Nouvel estimateur de l'erreur-type pour les groupes aléatoires

WILLARD C. LOSINGER[1]

RÉSUMÉ

La méthode utilisée pour estimer l'erreur-type des données des recensements décennaux des États-Unis de 1970 à 1990 donne des résultats disparates. Ainsi, on obtient des erreurs-types différentes pour les réponses «oui» et «non» à la même variable binomiale, alors que les deux estimations devraient être identiques. Quand la plupart des personnes répondent d'une façon à une question binomiale et quelques autres donnent la réponse contraire, l'erreur-type est beaucoup plus élevée pour la réponse la plus fréquente. D'autre part, lorsque les personnes interrogées fournissent toutes la même réponse, l'erreur-type n'est pas égale à zéro et reste fort élevée. Signaler les effets moyens du plan d'échantillonnage pondérés selon le nombre de répondants qui présentent des caractéristiques particulières ne fait qu'aggraver le problème. L'auteur propose une solution de rechange à la méthode d'estimation de l'erreur-type des groupes aléatoires utilisée dans le cadre du recensement des États-Unis.

MOTS CLÉS: Recensement; estimation de la variance; groupes aléatoires; effet du plan d'échantillonnage.

1. INTRODUCTION

Lors du recensement décennal de 1990, tous les répondants ont été priés de fournir des renseignements sur certains éléments de donnée (baptisés «données intégrales»). La plupart des personnes ont inscrit l'information requise sur le questionnaire abrégé du recensement. Par ailleurs, dans le cadre d'un échantillonnage systématique (allant du huitième à la demie, mais se situant en moyenne au sixième), on a demandé aux répondants des précisions sur d'autres éléments de donnée (données d'échantillon), au moyen du questionnaire détaillé du recensement.

Au lieu d'estimer l'erreur-type de toutes les données échantillonnées diffusées, le Census Bureau a préparé des tableaux indiquant les effets généralisés du plan d'échantillonnage. L'utilisateur devait calculer l'erreur-type en supposant un échantillonnage aléatoire simple (soit avec la formule normalisée, soit à partir d'un tableau) et un taux d'échantillonnage d'une personne sur six pour toutes les données d'échantillon. Il suffisait ensuite de multiplier le résultat obtenu par l'effet généralisé du plan d'échantillonnage (présenté dans un autre tableau). Le tableau en question énumérait les effets généralisés du plan d'échantillonnage sur l'élément de donnée concerné et la proportion de personnes ou d'unités de logements dans l'échantillon (on trouvera au tableau 1 les effets du plan d'échantillonnage diffusés pour les données d'échantillon du Vermont en 1990). Ainsi, pour toutes les estimations se rapportant à la profession qui ont été rendues publiques, l'utilisateur a pu constater que le plan d'échantillonnage avait quatre effets généralisés, soit un pour chaque taux d'échantillonnage applicable aux personnes, dans le rapport. Pour connaître l'erreur-type relative au nombre d'enseignants dans un rapport, l'utilisateur devait multiplier l'erreur-type de l'échantillonnage aléatoire simple (en supposant un taux d'échantillonnage d'une personne sur six, selon la formule ou d'après le tableau des erreurs-types) par l'effet du plan d'échantillonnage sur les éléments de donnée se rapportant à la profession qui correspondait au taux d'échantillonnage indiqué. Le nombre estimatif d'enseignants et l'erreur-type permettaient ensuite d'établir un intervalle de confiance. On trouvera d'autres explications sur l'usage du tableau des effets du plan d'échantillonnage à la partie «Accuracy of the Data» des produits de données (U.S. Bureau of the Census 1993, par exemple).

2. ESTIMATION DE L'ERREUR-TYPE

On s'est servi de la méthode des groupes aléatoires pour estimer l'erreur-type des données échantillonnées lors du recensement. Les États-Unis ont été divisés en un peu plus de 60,000 zones distinctes baptisées «régions de pondération» pour lesquelles on a calculé un poids d'échantillonnage. Dans chaque région de pondération, les unités d'échantillonnage (unités de logement ou personnes vivant dans un logement collectif) ont été systématiquement réparties entre 25 groupes aléatoires. On pensait qu'ainsi chaque groupe aléatoire respecterait à peu près le même plan d'échantillonnage que l'échantillon original (Wolter 1985).

On a estimé séparément le total de chacun des 1,804 éléments de donnée d'échantillon pour chaque groupe aléatoire en multipliant la population pondérée de l'élément de donnée échantillonné du groupe par 25. L'erreur-type des groupes aléatoires a ensuite été déterminée pour chaque élément de donnée dont le nombre total de personnes présentant un caractéristique particulière avait été estimé à partir des données échantillonnées au moyen des 25 estimations différentes du total venant des groupes aléatoires:

$$S_{\mathrm{RG}} = \sqrt{(1 - n/N)\sum_{i=1}^{25} \frac{(\hat{Y}_i - \hat{Y})^2}{24}}$$

[1] Willard C. Losinger, U.S. Department of Agriculture: APHIS:VS, CEAH, 555 South Howes Street, Suite 200, Fort Collins, CO 80521, U.S.A.

BIBLIOGRAPHIE

JUDKINS, D., MASSEY, J., et WAKSBERG, J. (1992). Patterns of residential concentration by race and Hispanic origin. *Proceedings of the Social Statistics Section, American Statistical Association,* 51-60.

KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A,* 149, 1, 65-82.

KISH, L. (1965). *Survey Sampling.* New York: Wiley.

MASSEY, J., JUDKINS, D., et WAKSBERG. J. (1993). Collecting health data on minority populations in a national survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 75-84.

UNITED NATIONS (1993). *Sampling Rare and Elusive Populations.* Department for Economic and Social Information and Policy Analysis, Statistical Division, National Household Survey Capability Programme. New York.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 429-434.

WAKSBERG, J. (1995). Distribution of poverty in Census block groups (BG's) and implications for sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 497-502.

WAKSBERG, J., et MOHADJER, L. (1991). Automation of within-household sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 350-355.

Tableau 7
Mélange résidentiel des minorités

Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de Noirs vivant dans la strate en 1990			Pourcentage de personnes d'ascendance hispanique vivant dans la strate en 1990			Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990			Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		
	Domaine de stratification			Domaine de stratification			Domaine de stratification			Domaine de stratification		
	Hispaniques	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux, et Aléoutes	Noirs	Natifs d'Asie et du Pacifique	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Amérindiens, Esquimaux et Aléoutes	Noirs	Hispaniques	Natifs d'Asie et du Pacifique
< 10%	79.2	95.4	99.6	73.4	86.3	99.1	78.9	59.2	99.6	85.9	81.4	95.1
10-30%	12.7	3.8	0.3	15.5	10.7	0.8	15.2	26.9	0.4	8.2	12.3	3.9
30-60%	5.8	0.7	0.0	7.4	2.5	0.1	4.2	10.8	0.0	3.3	4.5	0.8
60-100%	2.2	0.1	0.0	3.6	0.5	0.1	1.6	3.2	0.0	2.5	1.8	0.2

Sources: Recensement décennal de 1990 (totalisation Westat)

natifs d'Asie et du Pacifique représentent plus de 10% de la population n'est que de 13.7%, alors que le pourcentage de natifs d'Asie et du Pacifique dans les îlots où la population de personnes d'origine hispanique dépasse 10% de la population s'élève à 40.8%. Sur le plan pratique, ce chevauchement présente sans doute peu d'importance, on aurait besoin d'un échantillon de présélection très élevé (dans les régions très concentrées et à l'extérieur de celles-ci) pour trouver un nombre suffisant de natifs d'Asie et du Pacifique en vue de respecter les besoins de précision modérés, pareil échantillon ferait ressortir un nombre suffisant de personnes d'origine hispanique sans qu'on doive recourir à une répartition disproportionnée de l'échantillon entre les îlots où se concentrent davantage les personnes d'ascendance hispanique.

11. CONCLUSIONS

Le suréchantillonnage géographique reposant sur les données issues du recensement décennal le plus récent est une stratégie utile lorsqu'on veut accroître la précision des statistiques sur la population de race noire et d'origine hispanique dans les enquêtes-ménages entreprises aux États-Unis, pourvu que le coût de l'interview complète soit de 5 à 10 fois moindre que celui de l'interview de présélection. La même stratégie donne aussi de bons résultats lorsqu'il s'agit d'améliorer la précision des statistiques sur les natifs d'Asie et du Pacifique ou sur les Amérindiens, les Esquimaux et les Aléoutes, même lorsque le rapport entre le coût de l'interview intégrale et le coût de l'interview de présélection est très élevé.

Quoi qu'il en soit, ces constatations ne signifient pas qu'on puisse réaliser une enquête à un coût raisonnable en vue d'obtenir simultanément des statistiques très précises sur toutes les sous-populations désirées et de garder le degré de précision souhaité pour la population globale. La majorité des enquêtes démographiques réclament une précision raisonnable pour les domaines spécialisés et la population, en général. En déplaçant une portion de l'interview intégrale de la population blanche d'ascendance non hispanique vers les autres sous-populations, on devrait réduire la précision des statistiques sur la population totale. Il vaut habituellement la peine de parvenir à un compromis entre la précision des données sur les sous-populations et sur la population totale. Par cela, nous voulons simplement faire ressortir que le suréchantillonnage géographique ne met pas fin à la nécessité de sélectionner de très gros échantillons et d'entreprendre de nombreuses interviews de présélection quand on s'efforce de recueillir des statistiques précises sur des domaines rares, au coût le plus bas. Les statistiques précises sur des sous-populations rares continueront d'être dispendieuses qu'on recoure ou non au suréchantillonnage géographique.

En ce qui concerne les enquêtes sur les défavorisés, le suréchantillonnage géographique ne donne lieu qu'à de légers gains, et cela uniquement quand le coût de l'interview complète dépasse de plusieurs fois celui de l'interview de présélection et de l'abandon d'un ménage. Les gains devraient pour la plupart disparaître avec la détérioration due au passage du temps. En réalité, vers le milieu de la décennie ou un peu plus tard, une fois que les données du recensement seront sérieusement périmées, il se peut fort bien que le suréchantillonnage géographique entraîne une perte d'efficacité au lieu d'une amélioration, à cause de la migration des démunis et de l'erreur d'échantillonnage qui résulte de la quantification de la pauvreté au niveau des groupes d'îlots. Néanmoins, le suréchantillonnage géographique reste un instrument utile quand on s'intéresse surtout aux personnes défavorisées de race noire ou d'origine hispanique.

REMERCIEMENTS

Ce projet de recherche a été entrepris par Westat Inc. dans le cadre de l'entente 200-89-7021 parrainée par le National Center for Health Statistics et les Centers for Disease Control and Prevention. David Judkins et James Massey ont prêté leur concours au projet pendant la période où ils travaillaient respectivement à Westat et au NCHS. Les auteurs tiennent à remercier John Edmonds et Robert Dymowski, de Westat, pour leur participation à la programmation, ainsi que les arbitres pour leurs commentaires et observations sur une version antérieure de cet article.

Tableau 6

Ségrégation résidentielle des démunis selon la race et l'ascendance

Strate de densité (taux de pauvreté des personnes de la race ou de l'ascendance indiquée dans le groupe d'îlots en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990			Strate de densité (pourcentage de la minorité indiquée dans l'îlot en 1990)	Pourcentage de personnes de la race ou de l'ascendance indiquée dont le revenu était inférieur au seuil de pauvreté et qui vivaient dans la strate en 1990		
	Domaine				Domaine		
	Noirs	Hispaniques	Autres		Noirs	Hispaniques	Autres
< 5%	0.6	0.6	10.4	< 5%	4.0	4.6	n/a
5-10%	2.2	2.4	19.6	5-10%	3.7	5.1	n/a
10-20%	8.8	11.0	32.6	10-30%	13.2	19.9	n/a
20-30%	13.8	17.0	18.1	30-60%	19.0	25.5	n/a
30-40%	17.0	19.3	9.0	60-100%	60.0	44.8	n/a
40-50%	17.3	17.7	4.6				
50-100%	40.4	32.0	5.6				
Population totale (milliers)	8,557	5,536	17,975	Population totale (milliers)	8,557	5,536	17,975

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

L'analyse de tableaux plus détaillés (non présentés) révèle que l'efficacité est à peu près la même pour différentes ventilations géographiques, c'est-à-dire par État, par petite ou grande MSA, par noyau urbain, par banlieue ou par région non métropolitaine. Les conclusions tirées de cette analyse s'appliqueront donc approximativement aux enquêtes infranationales.

Le suréchantillonnage géographique demeure toutefois un instrument d'une très grande efficacité pour les populations noires et hispaniques à faible revenu. Comme on peut le constater au tableau 6, les personnes de race noire et celles d'ascendance hispanique qui vivent dans la pauvreté se regroupent beaucoup, à l'inverse des démunis des autres races. Le côté gauche du tableau 6 donne la distribution des défavorisés de race noire, d'origine hispanique et des autres groupes entre les strates de densité définies selon l'indice de pauvreté par rapport à la sous-population sur laquelle on se penche. En prenant un exemple du côté gauche, on se rend compte qu'en 1990, 32% des défavorisés d'origine hispanique vivaient dans un groupe d'îlots où l'indice de pauvreté pour les membres de ce groupe dépassait 50%. Le côté droit présente la distribution des défavorisés de race noire et d'ascendance hispanique pour les strates de densité définies d'après la concentration locale de noirs ou de personnes d'origine hispanique, sans égard au revenu. En prenant un exemple du côté droit, on note qu'en 1990, 44.8% des démunis d'origine hispanique habitaient un groupe d'îlots où les gens de la même ascendance représentaient plus de 60% de la population. Ces chiffres laissent entendre que plus de 90% des défavorisés de race noire et d'origine hispanique vivaient à un endroit où la concentration du groupe racial ou ethnique pertinent dépassait la moyenne. Bref, une stratégie prévoyant le suréchantillonnage des îlots à forte population de personnes de race noire ou d'ascendance hispanique débouchera automatiquement sur un nombre disproportionné de défavorisés. Par ailleurs, on ne retrouve presque aucun démuni de race noire ou d'origine hispanique dans les régions où ces deux groupes indiquent un faible indice de pauvreté. Il s'agit d'un contraste marqué avec les tendances relevées pour les défavorisés qui ne sont ni de race noire, ni d'ascendance hispanique. Apparemment, beaucoup de défavorisés de race blanche mais d'origine non hispanique vivent en voisinage avec d'autres personnes de race blanche mais nanties, peut-être parce que pour elles, la pauvreté a tendance à être un phénomène transitoire ou parce qu'il s'agit de personnes à la retraite qui avaient acheté leur maison à un moment où elles se trouvaient dans une meilleure situation.

10. SURÉCHANTILLONNAGE SIMULTANÉ DE PLUSIEURS DOMAINES RACIAUX-ETHNIQUES

En général, le suréchantillonnage géographique s'avère aussi facile et aussi efficace pour plusieurs domaines raciaux-ethniques que pour un seul. De fait, les taux d'échantillonnage optimaux pour les strates où se concentre chaque sous-population à laquelle on s'intéresse sont à peu près les mêmes que si on se penchait sur un seul domaine. Le taux de présélection global augmente néanmoins, car le nombre de régions à taux d'échantillonnage élevé augmentera avec le nombre de domaines. Ces deux observations résultent du chevauchement restreint entre les régions où il y a forte ségrégation des minorités raciales et ethniques à l'étude.

Le tableau 7 fournit quelques données sur la question, du recensement décennal de 1990. Les seules sous-populations pour lesquelles on remarque un chevauchement important dans les régions à forte concentration sont les personnes d'ascendance hispanique et les natifs d'Asie et du Pacifique. Toutefois, même dans ce cas, le chevauchement ne s'effectue que dans un sens. Comme on compte beaucoup plus de personnes d'origine hispanique que de ressortissants d'Asie et des îles du Pacifique aux États-Unis, la proportion de personnes d'ascendance hispanique habitant les îlots où les

signifie que le suréchantillonnage des ménages de la strate où on note un pourcentage relativement élevé de personnes à faible revenu ne donnera pas de résultats vraiment supérieurs au suréchantillonnage et à la présélection de la base de sondage complète, à moins que le coût de l'interview intégrale ne dépasse que légèrement celui de l'interview de présélection.

La figure 5 indique le ratio de la variance entre l'échantillon optimal et un échantillon aléatoire simple du même coût, pour les statistiques se rapportant aux défavorisés. Fait intéressant, en dépit de la plus forte concentration qu'entraîne la définition la plus large de «faible revenu», la réduction de la variance attribuable au suréchantillonnage géographique est plus importante avec la définition la plus étroite, car celle-ci exige une présélection accrue, si bien qu'on a plus à gagner d'une stratégie d'échantillonnage qui atténue la présélection. Dans les trois cas, le suréchantillonnage semble déboucher sur des avantages modérés quand c est inférieur à 3 ou 4, soit une réduction de 10% ou de 15% de la variance. Lorsque c atteint 10, les gains sont minimes et le suréchantillonnage des GI n'offre virtuellement aucun avantage avec une concentration élevée de démunis, pour les valeurs de c égales ou supérieures à 20. Bien sûr, il faut aussi tenir compte de la migration, mais nous n'avons pu obtenir les données voulues. La migration devrait se traduire par une réduction de la variance presque certainement plus faible que celle indiquée au tableau. En outre, les données sur le revenu du recensement de 1990 reposent sur un échantillon d'un sixième. La taille de l'échantillon dans un groupe d'îlots typique est légèrement inférieure à 100 ménages. La classification des îlots d'après la proportion de personnes à faible revenu entraîne donc un degré de confusion appréciable et beaucoup de groupes d'îlots ne se retrouveront pas dans la catégorie que leur attribuent les données du recensement, mais dans une classe voisine, ce qui atténuera encore plus la réduction de la variance que l'on peut espérer avec le suréchantillonnage géographique. Pour ces raisons, il est peu probable que le suréchantillonnage géographique débouche sur des gains d'efficacité. En réalité, la variance pourrait augmenter vers le milieu de la décennie ou plus tard. Une étude connexe, inédite, effectuée par Waksberg en 1989 aboutit à des résultats similaires quand on envisage de fusionner les données sommaires sur le revenu selon le code zip aux banques de numéros de téléphone utilisées lors de l'échantillonnage par composition aléatoire. Les gains réalisables grâce à la stratification semblent fort limités.

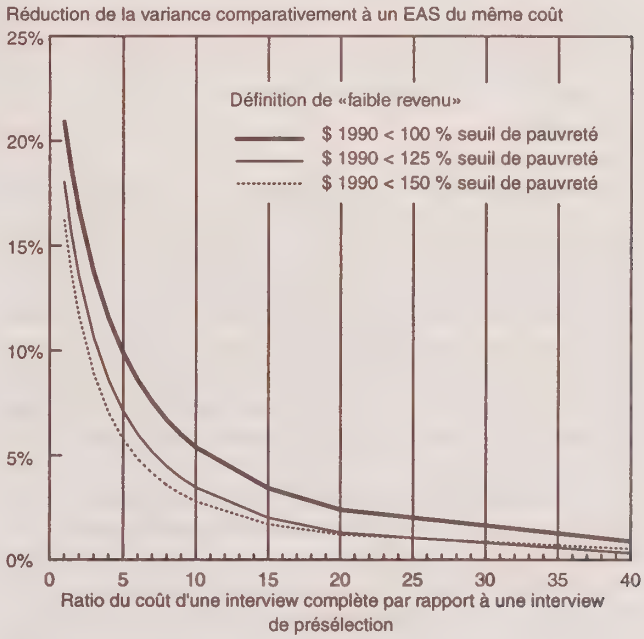


Figure 5. Réduction de la variance par suréchantillonnage géographique des démunis

Tableau 5
Ségrégation résidentielle des démunis

Strate de densité (pourcentage de démunis dans le groupe d'îlots en 1990 selon différentes définitions de «faible revenu»)	Pourcentage des démunis vivant dans la strate en 1990			Pourcentage de la population totale vivant dans la strate en 1990		
Définition de «faible revenu»:	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté	$ < pauvreté	$ < 125% pauvreté	$ < 150% pauvreté
< 5%	5.8	3.2	1.8	33.3	22.4	15.4
5-10%	12.3	8.3	5.7	22.1	19.7	16.7
10-20%	24.8	21 .0	16.8	22.8	25.2	24.8
20-30%	19.8	20.2	19.2	10.7	14.4	16.8
30-40%	14.3	15.9	17.0	5.4	8.1	10.7
40-50%	10.0	12.2	13.7	2.9	4.8	6.7
50-100%	13.0	19.3	25.7	2.8	5.4	8.8
Population totale (milliers)	31,797	42,316	52,521	248,710	248,710	248,710
Pourcentage de démunis aux États-Unis l'année de mesure	12.8	17.0	21.1			

Source: Recensement décennal de 1990 (totalisation Westat de STF-3)

Tableau 3
Ségrégation résidentielle des natifs d'Asie et du Pacifique

Strate de densité (pourcentage de natifs d'Asie et du Pacifique dans l'îlot ou le groupe d'îlots en 1990)	Pourcentage de natifs d'Asie et du Pacifique vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité de stratification:	GI	Îlot	GI	Îlot
< 5%	30.5	19.4	86.4	85.2
5-10%	17.2	17.7	7.2	7.4
10-30%	27.8	32.1	5.0	5.7
30-60%	14.6	18.0	1.0	1.3
60-100%	9.8	13.0	0.4	0.5
Population totale (milliers)	6,968	6,968	248,710	248,710
Pourcentage de natifs d'Asie et du Pacifique aux États-Unis l'année de mesure	2.8	2.8		

Source: Recensement décennal de 1990 (totalisation Westat)

Tableau 4
Ségrégation résidentielle des Amérindiens, des Esquimaux et des Aléoutes

Strate de densité (pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes dans l'unité de stratification en 1990)	Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes vivant dans la strate en 1990		Pourcentage de la population totale vivant dans la strate en 1990	
Unité stratification:	GI	Îlot	GI	Îlot
< 5%	50.3	34.6	98.3	97.4
5-10%	7.4	12.1	0.8	1.4
10-30%	12.4	15.9	0.6	0.8
30-60%	6.0	7.7	0.1	0.1
60-100%	23.8	29.6	0.2	0.2
Population totale (milliers)	1,793	1,793	248,710	248,710
Pourcentage d'Amérindiens, d'Esquimaux et d'Aléoutes aux États-Unis l'année de mesure	0.7	0.7		

Source: Recensement décennal de 1990 (totalisation Westat)

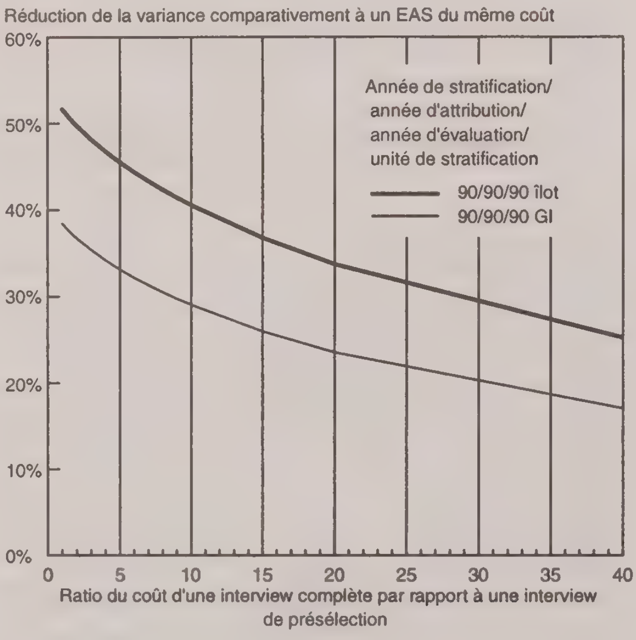


Figure 4. Réduction de la variance par suréchantillonnage géographique des Amérindiens, des Esquimaux et des Aléoutes

9. SURÉCHANTILLONNAGE DES DÉMUNIS

Le tableau 5 montre la distribution de la population défavorisée par groupe d'îlots, classée selon la population de personnes à faible revenu dans les GI, en 1990. Le nombre de GI dans chaque classe varie avec la définition de «faible revenu». Les chiffres qui apparaissent dans le tableau correspondent au pourcentage de personnes défavorisées dans chaque classe. Le tableau 5 illustre une distribution relativement uniforme de la pauvreté entre les différentes classes, pour les trois définitions retenues en 1990. Les données (non indiquées) du recensement de 1970 et du Current Population Survey révèlent que la ségrégation des personnes sous le seuil de pauvreté s'est accrue entre 1970 et 1990 (Waksberg 1995), mais elle reste beaucoup moins importante que la ségrégation des groupes ethniques et raciaux. La concentration est légèrement plus élevée pour le groupe de personnes sous la marque de 150% que pour les groupes nés des deux autres définitions mais, même pour ce groupe, elle demeure considérablement inférieure à celle observée pour les groupes raciaux et ethniques. Comme on peut le constater, avec cette définition, 25% seulement environ des démunis vivent dans des GI comptant 50% ou plus de démunis. Les pourcentages correspondants sont de 19% pour les personnes sous la marque de 125% de pauvreté et de 13% seulement pour celles sous la marque de 100% de pauvreté. Pareille distribution

d'être mentionnés. D'abord il semble qu'à l'inverse des noirs, la ségrégation ait légèrement augmenté pour les membres de ce groupe, entre 1980 et 1990. Néanmoins, il existe des tendances analogues pour les deux populations. Ainsi, en 1980, 30% de la population d'origine hispanique vivait dans des groupes d'îlots comprenant au moins 60% de membres du même groupe. En 1988, ces groupes d'îlots ne comprenaient plus qu'environ 21% de personnes d'ascendance hispanique. En revanche, la proportion de personnes d'origine hispanique habitant un groupe d'îlots qui comptait moins de 5% de personnes de même ascendance en 1980 est passée de 15% à 29% entre 1980 et 1988. Cette évolution traduit un déplacement des personnes d'origine hispanique entre les régions, et un accroissement de la population hispanique aux États-Unis. La restratification de la population hispanique à partir des données de 1990 fait ressortir des tendances similaires à celles révélées par la distribution de 1980.

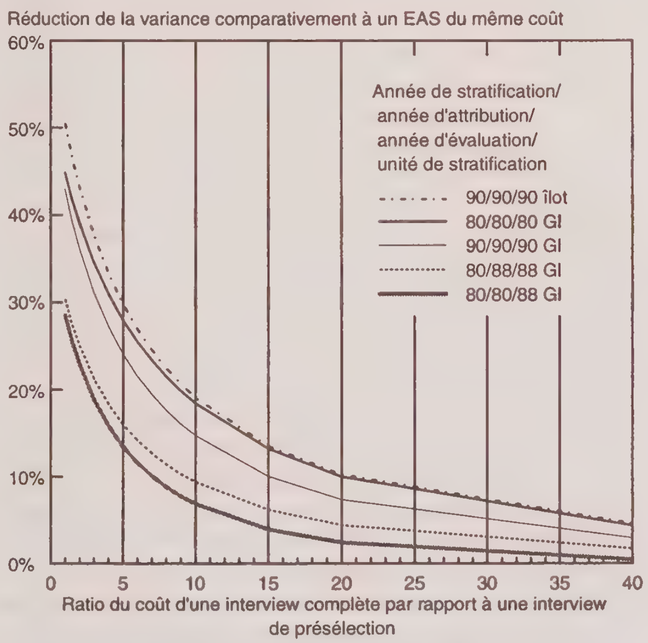


Figure 2. Réduction de la variance par suréchantillonnage géographique des personnes d'ascendance hispanique

La figure 2 donne un aperçu des répercussions de cette ségrégation sur les plans de suréchantillonnage. Les courbes suivent la même tendance générale que celles de la population noire et le suréchantillonnage géographique paraît utile pour les valeurs $c < 10$. Encore une fois cependant, il convient de se rappeler les effets de la migration sur la réduction de la variance. L'écart entre les lignes 80/80/80 et 80/80/88 est plus grand pour la population hispanique que pour la population noire, surtout aux valeurs $c < 5$. Pour l'instant, on ne dispose pas d'une base assez solide pour dire que la situation des années 80 se répétera dans les années 90.

8. SURÉCHANTILLONNAGE DES AUTRES MINORITÉS RACIALES

Les tableaux 3 et 4 présentent les données sur la ségrégation des natifs d'Asie et du Pacifique, d'une part, et des Amérindiens, des Esquimaux et des Aléoutes, d'autre part. Les figures 3 et 4 décrivent les implications correspondantes sur le suréchantillonnage de ces minorités. Les données de 1980 et de 1988 n'ont pas été reportées au tableau pour cette analyse, car les données de 1990 ne laissent pas entrevoir un suréchantillonnage bon marché des populations concernées, même avec une stratification selon la densité de population. La réduction en pourcentage de la variance est fort importante; elle dépasse celle obtenue avec les populations noire et hispanique, car la présélection nécessaire est beaucoup plus importante. Toutefois, le fait qu'il s'agisse de très petites populations aux États-Unis signifie qu'on aura toujours besoin d'une vaste présélection pour disposer d'un échantillon de taille suffisante au moment de l'interview. Ainsi, avec un ratio de coût égal à trois, il faudrait sélectionner 61,000 personnes (soit environ 24,000 ménages) pour obtenir un échantillon d'Amérindiens, d'Esquimaux et d'Aléoutes ayant la même précision qu'un échantillon aléatoire simple (théorique) de 1,000 personnes, de cette population. (Bien sûr, pour présélectionner 24,000 ménages, on devrait choisir un nombre supérieur d'unités de logement, de manière à tenir compte des logements vacants et des non-réponses). Les chiffres correspondant pour les natifs d'Asie et du Pacifique sont de 18,000 personnes, ou environ 7,000 ménages.

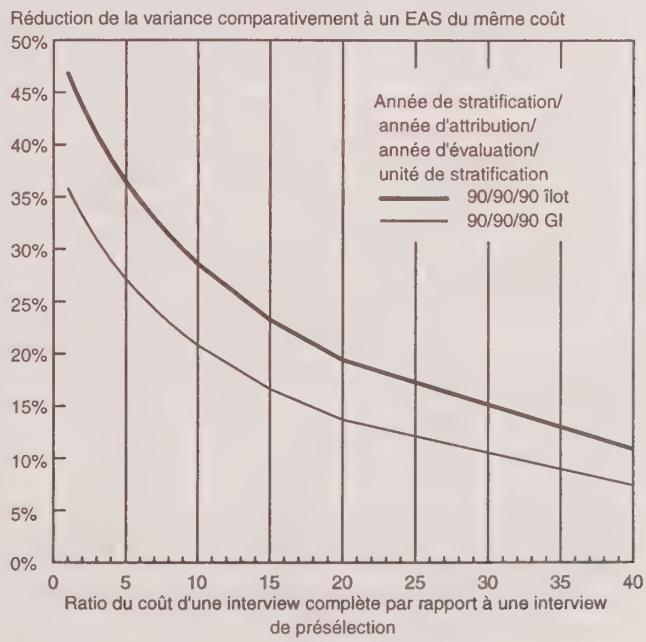


Figure 3. Réduction de la variance par suréchantillonnage géographique des natifs d'Asie et du Pacifique

Tableau 1
Ségrégation résidentielle de la population noire

Strate de densité (pourcentage de Noirs dans l'unité de stratification l'année de la stratification)	Pourcentage de Noirs vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 10%	9.7	20.5	12.0	8.5	78.2	81.4	75.7	77.5
10-30%	13.5	13.2	16.8	13.9	8.9	7.1	11.4	9.6
30-60%	18.9	20.4	20.3	16.2	5.1	5.1	5.7	4.5
60-100%	57.9	45.9	51.0	61.4	7.8	6.4	7.2	8.4
Population totale (milliers)	26,495	29,380	29.986	29,986	226,546	240,876	248,710	248,710
Pourcentage de Noirs aux États-Unis à l'année de la mesure	11.7	12.0	12.1	12.1				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

d'îlots où existait une forte concentration de population noire en 1980 pour d'autres groupes d'îlots que la réduction de la variance réalisable par suréchantillonnage avait presque diminué de moitié. La migration des noirs vers des groupes d'îlots où les gens de cette race étaient moins nombreux en 1980 augmente le nombre de noirs de l'échantillon auxquels est associé un facteur de pondération important, ce qui accroît la variabilité des poids, donc la variance. Quoi qu'il en soit, la réduction de la variance qu'indique la ligne 80/80/88 pour les valeurs $c < 10$ suffit indubitablement pour être jugée utile.

Lorsqu'on examine les données de 1990 à la figure 1, on constate que la ligne 90/90/90 des GI reste toujours quelques points sous la ligne 80/80/80, signe que le suréchantillonnage géographique des groupes d'îlots est un peu moins utile dans les années 90 que dans les années 80. On le doit à la légère diminution de la ségrégation des Américains de race noire en 1990, comparativement à la situation qui prévalait en 1980, ainsi qu'on l'a déjà souligné. D'un autre côté, la ligne 90/90/90 pour les ilôts est presque identique à la ligne 80/80/80. Le suréchantillonnage géographique des îlots devrait donc donner d'aussi bons résultats dans les années 90 que le suréchantillonnage des groupes d'îlots dans les années 80. Bien qu'on n'ait pas encore recueilli de données sur la distribution de la population noire à la fin des années 90 entre les strates de densité de 1990, il est raisonnable de croire que la migration s'est poursuivie, bref que les gains signalés par les lignes de 1990 soient plus faibles (conformément à la tendance générale que révèle la ligne 80/80/88), quand on établit des prévisions sur les économies envisageables pour la fin des années 90 et les premières années qui suivront l'an 2000.

7. SURÉCHANTILLONNAGE DES PERSONNES D'ORIGINE HISPANIQUE

Le tableau 2 présente divers aspects de la ségrégation résidentielle des personnes d'ascendance hispanique aux États-Unis qu'il est important de connaître lorsqu'on élabore une enquête sur la population. Plusieurs points valent la peine

Tableau 2
Ségrégation résidentielle des personnes d'ascendance hispanique

Strate de densité (pourcentage de personnes d'ascendance hispanique dans l'unité de stratification l'année de la stratification)	Pourcentage de personnes d'ascendance hispanique vivant dans la strate l'année indiquée				Pourcentage de la population totale vivant dans la strate l'année indiquée			
Année de mesure	1980	1988	1990	1990	1980	1988	1990	1990
Année de stratification	1980	1980	1990	1990	1980	1980	1990	1990
Unité de stratification	GI/DD	GI/DD	GI	Îlot	GI/DD	GI/DD	GI	Îlot
< 5%	14.8	29.3	10.6	6.6	76.8	79.8	68.4	68.9
5-10%	9.6	9.5	8.7	8.1	8.8	7.7	10.9	10.3
10-30%	22.6	21.2	22.8	22.1	8.5	7.4	11.8	11.5
30-60%	23.1	18.8	24.1	23.3	3.5	3.0	5.1	4.9
60-100%	30.0	21.2	33.9	39.8	2.4	2.0	3.8	4.4
Population totale (milliers)	14,609	19,393	22,354	22,354	226,546	240,876	248,710	248,710
Pourcentage de personnes d'ascendance hispanique aux États-Unis l'année de mesure	6.4	8.1	9.0	9.0				

Sources: Recensement décennal de 1980 (totalisation Westat)
Enquête nationale sur la santé de 1988 (totalisation Westat)
Recensement décennal de 1990 (totalisation Westat)

du Pacifique, les Amérindiens, les Esquimaux, les Aléoutes et les personnes à faible revenu, mais on ne l'a pas fait.

Les tableaux et les graphiques qu'on trouvera dans le reste de l'article présentent les données de diverses sources à divers moments dans le temps. Il convient de se rappeler que les données dont on s'est servi pour créer les strates ne sont pas nécessairement identiques à celles utilisées pour répartir l'échantillon. Les données qui ont permis d'évaluer l'échantillon peuvent venir d'un troisième point dans le temps ou d'une troisième source. Le présent article examine les combinaisons suivantes:

Étiquette	Source des données (stratification)	Source des données (répartition)	Source des données (évaluation)
80/80/80 GI	Recensement de 1980 (GI)	Recensement de 1980	Recensement de 1980
80/80/88 GI	Recensement de 1980 (GI)	Recensement de 1980	ENS 1988
80/88/88 GI	Recensement de 1980 (GI)	ENS 1988	ENS 1988
90/90/90 GI	Recensement de 1990 (GI)	Recensement de 1990	Recensement de 1990
90/90/90 îlot	Recensement de 1990 (îlot)	Recensement de 1990	Recensement de 1990

6. SURÉCHANTILLONNAGE DE LA POPULATION NOIRE

Le tableau 1 présente différents aspects de la ségrégation résidentielle de la population noire aux États-Unis, aspects importants à connaître lorsqu'on entreprend une enquête sur la population. Bien que la proportion de noirs habitant dans des groupes d'îlots à forte concentration de population noire (60% et plus) ait diminué entre 1980 et 1990, la ségrégation demeure très forte au sein de ce groupe. Les colonnes indiquant la population en 1988 revêtent une importance particulière, car elles illustrent la dynamique des données stratifiées dans le temps. En 1988, la proportion de noirs qui vivaient dans un groupe d'îlots où l'on comptait moins de 10% de noirs en 1980 avait doublé, de 9.7% à 20.5%. Ce phénomène a d'importantes répercussions sur l'efficacité du suréchantillonnage géographique, comme on le verra plus loin. Il est aussi intéressant de noter que la population totale des groupes d'îlots à forte concentration de personnes de race noire (c'est-à-dire où l'on recensait plus de 60% de noirs) en 1980 a diminué d'environ 2 millions d'habitants entre 1980 et 1988. La tendance résulte donc en partie de l'abandon de quelques vieux logements ou quartiers. La concentration est plus élevée au niveau de l'îlot qu'à celui du groupe d'îlots, en 1990, ainsi qu'on s'y attendait. (Les données sur les îlots ne sont plus disponibles pour les États-Unis à partir de 1980.) Quoique l'échantillonnage d'îlots coûte légèrement plus cher que l'échantillonnage de groupes d'îlots (à cause du plus grand nombre d'îlots et de la nécessité de prendre certaines dispositions quand l'îlot compte moins d'habitants que la taille idéale de la grappe), elle permet de mieux cibler le domaine auquel on s'intéresse.

La figure 1 résume les effets des données sur la densité de population qui apparaissent au tableau 1 à l'égard du suréchantillonnage de la population noire. On constate l'effet sensible de c sur l'efficacité du suréchantillonnage géographique. Quand c dépasse 20, la meilleure façon d'échantillonner la population noire consiste vraisemblablement à sélectionner un échantillon équiprobable.

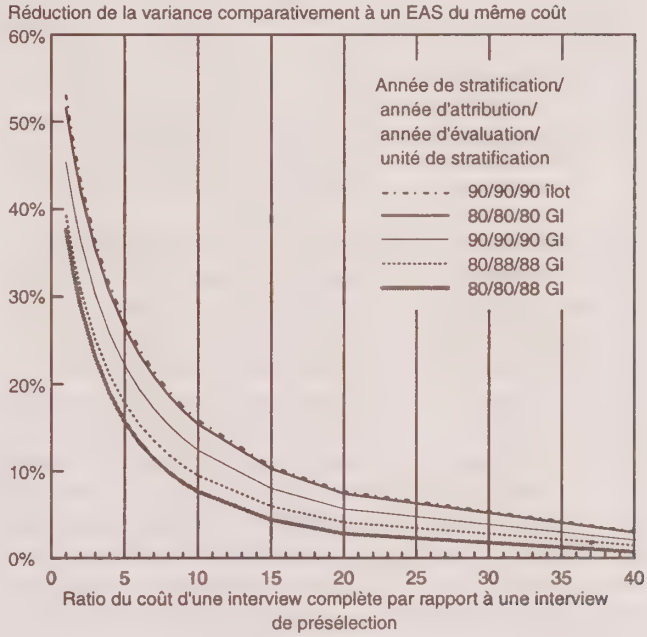


Figure 1. Réduction de la variance par suréchantillonnage géographique de la population noire

La figure montre aussi le risque qu'on court en ne se fiant qu'aux données stratifiées pour évaluer les avantages du suréchantillonnage géographique. La ligne 80/80/80 indique la réduction de la variance qu'on pourrait obtenir si la distribution de la population noire ne variait pas dans le temps selon les strates de densité définies au moyen des données sur les groupes d'îlots de 1980. La ligne 80/80/88 révèle la réduction réelle de la variance obtenue en 1988 grâce aux mêmes strates et à la même répartition. À $c = 5$, la variance diminue de 26% avec une distribution statique, alors qu'elle ne baisse que de 16% si la distribution change. Nous avons essayé de déterminer si la répartition de l'échantillon entre les anciennes strates, en fonction de nouvelles données sur la distribution réduirait la variance davantage en 1988. La réponse est oui, mais pas de beaucoup. La ligne 80/88/88 signale la réduction de la variance réalisable quand on applique la distribution de 1988 aux strates de 1980 et s'en sert pour une enquête en 1988. À $c = 5$, la nouvelle répartition réduit la variance de 18%, modeste amélioration par rapport à la diminution de 16% que permet la répartition selon l'ancienne distribution. Nous en avons conclu que la principale difficulté résidait dans l'ancienne stratification. En 1988, un si grand nombre de noirs avaient quitté les groupes

$$\frac{n_D}{deff} = \frac{NP}{\left(\sum_h A_h / f_h\right)}. \tag{7}$$

Si on remplace les formules (1) et (4) dans (7), on obtient

$$\frac{n_D}{deff} = \frac{n\left(c - 1 + \frac{1}{P}\right)}{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}. \tag{8}$$

Cette dernière formule nous permet de comparer la variance d'une statistique arbitraire sur le domaine D après suréchantillonnage géographique à la variance de la même statistique résultant d'un échantillonnage aléatoire simple pour le même coût total B. On peut récrire la formule (8) de façon algébrique afin d'établir la part de la variance de l'échantillonnage aléatoire simple supprimée par le suréchantillonnage géographique

$$\frac{\frac{\sigma^2}{n} - \frac{\sigma^2 deff}{n_D}}{\frac{\sigma^2}{n}} =$$

$$1 - \frac{\left(\sum_h A_h \sqrt{c - 1 + \frac{1}{P_h}}\right)\left(\sum_h \frac{N_h P_h}{NP} \sqrt{c - 1 + \frac{1}{P_h}}\right)}{\left(c - 1 + \frac{1}{P}\right)}. \tag{9}$$

Il peut sans aucun doute arriver que la réduction soit négative, bref que l'échantillonnage aléatoire simple ait une variance inférieure pour le même coût. Pareille situation survient le plus souvent avec une strate où $NPA_h >> N_h P_h$, bref une strate qu'on croyait contenir une très petite partie de D mais qui en inclut en réalité une forte proportion. Soulignons que si $P_h \equiv P$, il ne faut s'attendre à aucune réduction de la variance consécutivement au suréchantillonnage géographique. De plus, à mesure que c s'approche de l'infini pour une valeur P fixe (ce qui correspond à une présélection de moins en moins chère par rapport à l'interview intégrale), la réduction de la variance approche la valeur nulle. Étant donné les complications supplémentaires qu'entraîne un échantillon stratifié, on en déduit que pour une valeur c élevée et une valeur P moyenne, la personne qui conçoit le plan d'échantillonnage devrait envisager un échantillonnage aléatoire simple au lieu de la stratification. La valeur du suréchantillonnage augmente à mesure que P s'approche de zéro, que c approche de 1, et que D se concentre dans une strate. Lorsque la sous-population D à laquelle on s'intéresse se retrouve dans une seule strate, l'efficacité de l'échantillon s'accroît car il y a de moins en moins de membres de D dans les autres strates, où la différence est plus importante. La partie qui suit évalue de façon empirique la réduction potentielle de la variance attribuable au suréchantillonnage géographique dans diverses conditions, pour plusieurs domaines démographiques.

5. ÉVALUATION EMPIRIQUE

Il est très difficile d'évaluer l'équation (9) pour les domaines auxquels on s'intéresse. On peut extraire des données sur P_h des bandes sommaires du recensement décennal diffusées par le Bureau of the Census au niveau de l'îlot, du groupe d'îlots et du district de dénombrement. Il est donc possible de définir des strates raisonnables et d'évaluer les équations (1) à (4). En supposant que P_h est statique dans le temps, on pourrait aussi évaluer les autres équations. Néanmoins, les Américains ont tendance à déménager souvent, ce qui modifie la composition raciale et ethnique de nombreux îlots (Judkins, Massey et Waksberg 1992). Quand les membres de D déménagent dans une région où ils étaient peu nombreux avant, les avantages du suréchantillonnage géographique s'amenuisent. Puisqu'on ne souhaite pas surestimer les avantages de la méthode, nous en avons cherché une autre afin d'obtenir une estimation raisonnable de A_h à certains points dans le temps après le recensement. Le couplage des données d'îlots ou de GI entre deux recensements consécutifs pourrait constituer une solution, mais il est irréalisable. Jusqu'à présent, les îlots ont été définis et étiquetés indépendamment d'un recensement à l'autre, et on n'a pas essayé de conserver les définitions en prévision d'une analyse longitudinale. C'est pourquoi il faut recourir à d'autres sources pour estimer A_h.

Les microdonnées issues des enquêtes-ménages que poursuit présentement le Bureau of the Census s'avèrent une bonne source d'information sur A_h, pour analyser les avantages de l'échantillonnage géographique à l'égard des populations de race noire et d'origine hispanique. Plus précisément, nous nous sommes servis des données du 1988 National Health Interview Survey (NHIS) effectuée par interview. Des employés du Bureau of the Census nous ont préparé une bande spéciale donnant le code du groupe d'îlots ou du district de dénombrement de 1980 pour tous les ménages ou presque interviewés lors de l'enquête de 1988, pour les habitations construites avant 1980. (Celles construites dans les années 80 auraient été échantillonnées à partir des permis de construction plutôt que par échantillonnage aérolaire. Pour ces cas, les étiquettes des îlots et les groupes d'îlots ne sont pas attribuées aux logements échantillonnés à cause de difficultés techniques.) Les données de l'ENS de 1988 ont ensuite été appariées aux fichiers sommaires du recensement de 1980 par groupe d'îlots ou district de dénombrement, ce qui a permis la répartition des ménages de l'ENS en strates définies selon la concentration de personnes de race noire et d'origine hispanique en 1980. Des facteurs de pondération ont ensuite servis à estimer la distribution de différentes sous-populations dans les strates. (On a supposé que les habitations construites dans les années 80 se trouvaient dans la strate où les sous-populations rares comptaient le moins de représentants.) Des opérations analogues auraient pu être effectuées pour les natifs d'Asie ou

donne l'essentiel des gains quand on se limite à un nombre assez faible de strates.

4. RÉPARTITION OPTIMALE POUR UN DOMAINE

L'objectif consiste à adapter les formules générales pour la répartition optimale d'un échantillon stratifié, afin de les appliquer à la réduction de la variance attribuable au suréchantillonnage géographique. Le développement mathématique correspond essentiellement à celui de Kish (1965), selon la notation proposée par Kalton dans United Nations (1993). Supposons que la population soit divisée en un certain nombre de strates, comme nous l'avons vu précédemment. Soit N, la taille de la population totale et N_h, la taille de la population de la h-ième strate. Soit P_h, la proportion de la h-ième strate représentant les membres de D et P, la proportion globale de la population que constitue D. On peut se servir du recensement décennal le plus récent pour estimer P_h et P, ou recourir à un autre sondage important plus récent donnant les codes des îlots et/ou des GI pour chaque ménage ou particulier échantillonné, de sorte que le couplage avec le dernier recensement décennal permettra d'affecter le ménage ou le particulier à une strate.

Nous supposerons que c est constant d'une strate à l'autre, même si ce n'est pas toujours le cas. Ainsi, procéder à des interviews dans les îlots où l'on trouve une grande concentration d'Amérindiens, d'Esquimaux ou d'Aléoutes signifie presque toujours se rendre dans des endroits éloignés où le transport s'avère difficile. Toutefois, même l'estimation d'une moyenne nationale pour c soulève des difficultés dans la plupart des enquêtes. En général, on ne pourra donc pas obtenir d'estimations par strates.

On suppose aussi que la distribution de $\boldsymbol{Y}$ dans D est constante d'une strate à l'autre. Plus précisément, on présume que

$$E(\boldsymbol{Y} \mid D \text{ et } h) \equiv E(\boldsymbol{Y} \mid D) \quad \text{et que}$$

$$\operatorname{Var}(\boldsymbol{Y} \mid D \text{ et } h) \equiv \operatorname{Var}(\boldsymbol{Y} \mid D),$$

où la valeur et la variance attendues se rapportent à la population, pas au plan d'échantillonnage. Cette hypothèse laisse habituellement à désirer, mais puisque les éléments du vecteur des caractéristiques auxquelles on s'intéresse se comportent souvent différemment d'une strate à l'autre, il est inutile d'essayer d'être plus précis. Enfin, on suppose que les fractions d'échantillonnage sont assez petites dans toutes les strates pour qu'on néglige les facteurs de correction de la population finie.

Des hypothèses qui précèdent, il ressort que la fraction d'échantillonnage optimale de la h-ième strate de l'enquête dans laquelle on délaisse tous les éléments U-D sélectionnés est

$$f_h = k\sqrt{\frac{P_h}{P_h(c-1)+1}}, \tag{1}$$

où k représente une constante déterminée par les exigences de précision ou les contraintes budgétaires. (Voir une des sources précitées pour une preuve de (1). Cette règle est une application de la répartition de Neyman.) Si $c = 1$ (c'est-à-dire si la présélection coûte aussi cher que l'interview), on peut simplifier l'équation qui précède en $f_h \propto \sqrt{P_h}$, ce qui peut déboucher sur une répartition fort différente d'un échantillon équiprobable pour toutes les strates. Si la présélection est beaucoup moins onéreuse que l'interview (c'est-à-dire si $c >> 1$) et que D n'est pas extrêmement rare (à savoir P_h n'approche pas de zéro), cette relation donne des fraction d'échantillonnage assez uniformes ce qui se traduit par une répartition proportionnelle à la population.

Étant donné un budget fixe B, on calcule k grâce à l'équation

$$B = \sum_h N_h f_h c' \left[P_h c + (1 - P_h)\right]. \tag{2}$$

Pour obtenir un échantillon aléatoire simple n du domaine D, il suffit de sélectionner un échantillon de taille n/P, qui coûtera en tout

$$B = ncc' + \left(\frac{n}{P} - n\right)c'. \tag{3}$$

En établissant l'égalité des coûts, on peut calculer la constante de proportionnalité de (1), qui devient:

$$k = \frac{n\left(c - 1 + \dfrac{1}{P}\right)}{\sum_h N_h P_h \sqrt{c - 1 + \dfrac{1}{P_h}}}. \tag{4}$$

Pour calculer de façon réaliste les avantages d'une telle répartition, on doit admettre que les estimations de P_h utilisées pour établir la répartition seront quelque peu désuètes au moment où se déroulera l'enquête. Soit A_h, la proportion réelle de D dans la h-ième strate lors de l'échantillonnage et de la collecte des données. On suppose que P n'a pas changé, même si la distribution entre les strates varie selon A_h. En établissant $NP = N_D$ et $N_D A_h = N_{dh}$, on se rend compte que la taille réelle de l'échantillon, n_D, pour D correspond à

$$n_D = \sum_h NPA_h f_h. \tag{5}$$

Selon Kish (1965), la variance de cet échantillon sera plus élevée que celle d'un échantillon aléatoire simple de même taille, pour D. Le facteur d'extension de la variance ou l'effet du plan d'échantillonnage associé aux différents taux de sondage entre les strates est donné par l'équation bien connue

$$deff = \left(\sum_h A_h f_h\right)\left(\sum_h A_h \big/ f_h\right). \tag{6}$$

Par conséquent, la taille *utile* de l'échantillon associé au suréchantillonnage géographique sera

soit D, une petite sous-population présentant un intérêt particulier, par exemple les personnes de race noire que l'on ne peut distinguer du reste de U lors de l'échantillonnage. Soit Y, la valeur vectorielle des caractéristiques auxquelles on s'intéresse, par exemple le revenu annuel, la situation d'emploi et le nombre de consultations médicales de l'année antérieure. Certaines enquêtes n'ont d'autre objectif qu'estimer la distribution de Y dans D. Dans une enquête de ce genre, les éléments U-D identifiés au moment de la présélection des membres de U seront supprimés de l'échantillon. Un questionnaire général peu coûteux permet d'effectuer cette présélection et d'établir qui recevra le questionnaire intégral.

On désire parfois estimer la distribution de Y simultanément dans D et dans U. Dans ce cas, on gardera au moins quelques membres U-D découverts lors de l'interview de présélection afin de leur faire passer l'interview complète. Lorsqu'on recourt au suréchantillonnage géographique, l'échantillon initial comprendra un suréchantillon des membres de U-D habitant dans les régions à forte concentration de population D. Même lorsqu'on s'intéresse à l'univers U-D, on s'efforce habituellement d'éviter le suréchantillonnage de U-D là où la population D est très concentrée car la variation de la probabilité de sélection de U-D entraîne des effets exagérés du plan d'échantillonnage sur les statistiques relatives à U et à U-D. À cause des effets plus marqués du plan d'échantillonnage, la taille de l'échantillon supplémentaire de U-D n'entraîne souvent qu'une réduction négligeable de la variance des statistiques relatives à U-D. On préfère généralement utiliser les fonds supplémentaires destinés aux interviews additionnelles auprès des membres de la population U-D pour accroître la taille de l'échantillon initial.

Il est assez facile d'établir des méthodes de sous-échantillonnage qui aboutiront à l'obtention d'un échantillon équiprobable de U-D. Le sous-échantillonnage peut-être centralisé après la présélection, ou l'intervieweur peut s'en charger avant de laisser le ménage sélectionné, en se renseignant sur la composition du ménage. En effet, on a mis au point des techniques qui facilitent considérablement le sous-échantillonnage par l'intervieweur (Waksberg et Mohadjer 1991). Il n'est pas nécessaire d'apprendre à ce dernier comment effectuer un tirage aléatoire. Lors des sondages papier et crayon, l'intervieweur reçoit des instructions sur les aspects qui devront être abordés avec tel ou tel ménage avant l'interview, d'une maison à l'autre. Ces instructions sont randomisées centralement avant la présélection en vue de donner les taux de sondage souhaités. Avec l'IPAO, il est possible de programmer le sous-échantillonnage pour qu'il s'effectue automatiquement sur l'ordinateur portatif; c'est l'ordinateur qui signale à l'intervieweur quel ménage répondra au questionnaire complet et quel ménage sera rejeté consécutivement au sous-échantillonnage.

La décision de garder tous les membres de U-D échantillonnés ou de procéder à un sous-échantillonnage dépend de la taille relative de U et de U-D, de la précision des données qu'on requiert sur les deux populations et du coût relatif de l'interview intégrale et de l'interview de présélection. Soit $c\star$, le coût variable associé à l'échantillonnage d'un membre de U ainsi qu'à la collecte et au traitement de ses données et soit c', le coût variable associé à l'échantillonnage, à la sélection et à l'abandon d'un membre de U. L'équation $c = c\star/c'$ exprime le rapport entre le coût d'une interview complète et celui de l'interview de présélection. Lorsque c est beaucoup plus grand que 1, et que l'enquête porte sur U-D on devrait envisager le sous-échantillonnage même si, ce faisant, l'enquête gagne en complexité. L'interview intégrale étant par définition plus longue que l'interview de présélection, c devrait toujours être légèrement supérieur à 1. Avec les enquêtes par panel et les enquêtes longitudinales, il conviendrait d'inclure le coût des interviews subséquentes à $c\star$, de sorte que le coût de l'interview complète pourrait devenir un très grand multiple du coût de l'interview de présélection c'est-à-dire, $c >> 1$. La même remarque s'applique aux enquêtes qui nécessitent le prélèvement de spécimens, et des travaux de laboratoire onéreux, ainsi qu'aux enquêtes qui font appel à des spécialistes coûteux (un médecin par exemple) pour la collecte des données primaires. Pour ce genre d'enquête, nous recommanderions fortement de ne pas recourir seulement au suréchantillonnage géographique mais de le combiner à la présélection et au sous-échantillonnage. Quand l'enquête se fait de porte à porte et ne comporte qu'une interview donnée par un intervieweur moyen (en mesure de poser des questions et d'enregistrer les réponses mais pas de procéder à une évaluation technique ou anthropologique), c se situe souvent entre 3 et 5, valeur suffisante dans de nombreux cas pour justifier le recours au sous-échantillonnage de U-D dans les régions suréchantillonnées.

3. STRATIFICATION

Même s'il est impossible de distinguer D de U au moment de l'échantillonnage, on suppose qu'on possède certains renseignements sur la distribution de D et de U dans une série d'entités géographiques. Aux États-Unis, les îlots ou groupes d'îlots (GI) constituent des éléments géographiques naturels et le recensement décennal nous renseigne sur eux. (Avant le recensement de 1990, il n'y avait pas d'ilôts dans les régions rurales; pour le suréchantillonnage, on se servait d'unités plus importantes baptisées «secteurs de dénombrement».) Le Bureau of the Census diffuse largement les données sur la composition raciale et ethnique des îlots, ainsi que des données cartographiques qui permettent aux organismes de sondage d'identifier ces îlots lors des années subséquentes. Les données sur le revenu ne sont fournies qu'au niveau des GI.

Il est courant de stratifier les îlots ou les GI en fonction de la concentration locale de D. Ainsi, les îlots où D représente moins de 10% de la population de l'îlot pourraient constituer une strate. D'autres strates pourraient correspondre à 30% et à 60% de la population, ce qui donnerait un total de quatre strates. Le nombre optimal de strates ou de points de découpage n'a guère fait l'objet d'études empiriques. En général, le plan d'échantillonnage est d'autant plus efficace qu'il compte plus de strates, mais il arrive un moment où les complexités opérationnelles soulevées par un grand nombre de strates dépassent les gains d'efficacité. Une certaine sagesse remontant à Kish (1965), veut que la stratification

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 69-80
Statistique Canada

Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis

JOSEPH WAKSBERG, DAVID JUDKINS et JAMES T. MASSEY[1]

RÉSUMÉ

Aux États-Unis, les enquêtes démographiques polyvalentes comptent souvent parmi leurs principaux objectifs la production d'estimations de petites sous-populations définies, par exemple, selon la race, l'origine ethnique et le revenu. Le suréchantillonnage géographique est l'une des techniques fréquemment envisagées pour accroître la fiabilité des statistiques sur ces petites sous-populations (ou domaines), dans la mesure où l'information sur les îlots ou les groupes d'îlots du Bureau of the Census permet d'identifier les régions où se concentrent les sous-populations auxquelles on s'intéresse. Les auteurs passent en revue les questions relatives au suréchantillonnage géographique des régions, parallèlement à la présélection des ménages en vue d'améliorer la précision des estimations sur les petits domaines. Ils présentent les résultats d'une évaluation empirique de la réduction de la variance obtenue par suréchantillonnage géographique et évaluent la robustesse de l'efficacité de l'échantillonnage dans le temps, à mesure que les données servant à la stratification perdent de leur actualité. L'article aborde aussi la question du suréchantillonnage simultané de plusieurs petites sous-populations.

MOTS CLÉS: Échantillonnage; stratification; populations rares.

1. INTRODUCTION

Ceux qui parrainent un grand nombre de vastes enquêtes démographiques polyvalentes souhaitent des analyses distinctes sur des sous-populations définies selon la race, l'origine ethnique et le revenu. En général, les échantillons équiprobables ne sont pas suffisants pour permettre une analyse assez précise de ces domaines, si bien qu'on doit procéder à un suréchantillonnage quelconque. Cette exigence soulève des problèmes méthodologiques intéressants, car aucun registre de la population américaine n'autorise le prélèvement d'échantillons stratifiés en fonction des domaines désirés. Le Bureau of the Census garde bien des listes des logements qui identifient les sous-populations désirées, mais les chercheurs de l'extérieur n'y ont pas accès. Quand l'enquête exige une interview personnelle, les chercheurs externes sont donc contraints à recourir aux techniques d'échantillonnage aréolaire. Le Bureau lui-même procède parfois à un suréchantillonnage géographique, car ses listes ne sont mises à jour que tous les dix ans. Nous décrirons des méthodes de suréchantillonnage efficaces pour les domaines précités, dans le contexte de l'échantillonnage aréolaire.

Les données du recensement décennal américain sur la concentration de diverses caractéristiques démographiques sont largement diffusées pour de petites unités géographiques; ainsi, on possède des données sur la race et l'origine ethnique pour chaque îlot et des données sur le revenu pour chaque groupe d'îlots. (Par «îlot», on entend une zone indivisible, circonscrite des quatre côtés par une route. Les groupes d'îlots sont constitués de plusieurs îlots contigus.) On peut recourir à ces données pour améliorer à peu de frais la précision des statistiques sur des sous-populations rares en suréchantillonnant les îlots ou les groupes d'îlots dans lesquels se concentrent les membres des groupes en question et en abandonnant ou en sous-échantillonnant les personnes qui n'appartiennent pas aux groupes rares recherchés. C'est à Kish (1965, partie 4.5) qu'on doit la théorie générale applicable aux plans d'échantillonnage de ce genre. Waksberg (1973) a fait une présentation personnelle de cette théorie en l'illustrant d'exemples tirés du recensement de 1960. Kalton et Anderson (1986) ont donné d'autres exemples pertinents et analysé les méthodes connexes, ce qu'on retrouve également dans un article rédigé par Kalton dans United Nations (1993). Dans le présent article, nous étendrons les illustrations antérieures à d'autres domaines, actualiserons les résultats à 1990 et évaluerons de façon empirique la robustesse de ces méthodes dans le temps.

Nous commencerons par passer brièvement en revue les aspects associés à l'élimination et au sous-échantillonnage des personnes qui ne font pas partie des groupes visés. Ensuite, nous examinerons la théorie de la répartition optimale en vertu de laquelle les strates sont définies d'après la densité des populations rares, et nous appliquerons cette théorie à plusieurs populations rares. La partie principale de l'article consiste toutefois en une évaluation empirique de la réduction de la variance qui résulte du suréchantillonnage géographique de diverses minorités et populations rares, ainsi que de la robustesse de telles réductions dans le temps. Enfin, nous aborderons les problèmes particuliers que pose l'échantillonnage simultané de plusieurs populations rares, avant d'établir nos conclusions.

2. COÛT DE L'ENQUÊTE ET DÉCISION D'EFFECTUER UNE PRÉSÉLECTION

Soit U, l'univers visé, par exemple les particuliers ou les ménages pour lesquels on dispose d'une base de sondage, et

[1] Joseph Waksberg, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.; David Judkins, Research Triangle Institute, 5901-B Peachtree-Dunwoody Road, Suite 500, Atlanta, GA 30325, U.S.A.; James T. Massey, autrefois de Westat Inc., maintenant décédé.

ce genre, la méthode d'estimation des variables instrumentales peut donner de bons résultats.

REMERCIEMENTS

Les auteurs remercient Wayne Fuller de qui vient l'idée à la base de cet article. Les recherches n'auraient pu être menées à bien sans l'aide fournie par l'Economic and Social Research Council (subvention H519 25 5005) dans le cadre de son programme Analysis of Large and Complex Datasets.

BIBLIOGRAPHIE

ABOWD, J.M., et ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics,* 3, 254-283.

ANDERSON, T.W. (1959). Some scaling models and estimation procedures in the latent class model. Probability and Statistics, (Éd. U. Grenander). Stockholm: Wiksell and Almquist.

BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis*. London: Griffin.

BIEMER, P.P., GROVES, R.M., LYBERG, L.E., MATHIOWETZ, N.A., et SUDMAN, S. (1991). *Measurement Errors in Surveys* New York: Wiley.

BRITTON, M., et BIRCH, F. (1985). *1981 Census Post-Enumeration Survey.* London: Her Majesty's Stationery Office.

CHUA, T., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

DURBIN, J. (1954). Errors in variables. *Revue de l'Institut International de Statistique*, 22, 23-31.

EDLEFSEN, L.E., et JONES, S.D. (1984). Reference Guide to GAUSS. Applied Technical Systems.

FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys*. (Éds., Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. et Sudman, S.). New York: Wiley.

FULLER, W.A. (1987). *Measurement Error Models*. New York: Wiley.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage.

HOGUE, C.R., et FLAIM, P.O. (1986). Measuring gross flows in the labor force: an overview of a special conference. *Journal of Business and Economic Statistics*, 41, 111-21.

MADANSKY, A. (1960). Determinental methods in latent class analysis. *Psychometrika*, 25, 183-198.

MARQUIS, K.H., et MOORE, J.C. (1990). Measurement errors in the Survey of Income and Program Participation (SIPP): Program Reports. *Proceedings of the 1990 Annual Research Conference*. US Bureau of the Census, 721-745.

MEYER, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

REIERSOL, D. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, 9, 1-24.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Ch. 3) (Éds. Skinner, C.J., Holt, D., et Smith, T.M.F). Chichester: Wiley.

SKINNER, C.J., et TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 53, 391-405.

VAN DE POL, F., et DE LEEUW, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15, 118-141.

VAN DE POL, F., et LANGEHEINE, R. (1990). Mixed Markov latent class models. Dans *Sociological Methodology 1990*, (Éd. C.C. Clogg). Oxford: Basil Blackwell, 213-247.

VAN DE POL, F., LANGEHEINE, R., et DE JONG, W. (1991). PANMARK User Manual. Panel analysis using Markov chains. Version 2.2. Netherlands Central Bureau of Statistics.

réalité, l'erreur-type n'est pas beaucoup plus importante que celle de l'estimateur non rajusté. L'intervalle de confiance (de deux erreurs-types) ne chevauche plus l'intervalle correspondant de l'estimateur non rajusté des quatre paramètres.

Tableau 3
Estimation non rajustée et estimation VI selon les données de l'EPDR

Paramètre	Estimation non rajustée	Estimations VI: IV = possession d'une voiture	IV = emploi déphasé	IV = emploi déphasé (Désagrégé)
$pr(x = 1, y = 1)$	0.719	0.773	0.766	0.757
	(0.006)	(0.033)	(0.008)	(0.007)
$pr(x = 1, y = 2)$	0.055	0.011	0.017	0.025
	(0.003)	(0.020)	(0.005)	(0.003)
$pr(x = 2, y = 1)$	0.061	0.018	0.024	0.032
	(0.003)	(0.019)	(0.004)	(0.003)
$pr(x = 2, y = 2)$	0.166	0.198	0.193	0.186
	(0.005)	(0.027)	(0.007)	(0.006)

Nota: Erreurs-types en vertu d'une hypothèse multinomiale entre parenthèses. Désagrégation selon l'âge (4 groupes), le sexe et le degré de scolarité (2 groupes).

Ainsi qu'on l'a vu précédemment, on peut remettre en question l'hypothèse (A4) pour la variable correspondant à l'emploi déphasé. La version désagrégée de l'estimateur pose des hypothèses "plus faibles", n'exigeant seulement que (A4) se confirme à l'intérieur d'un sous-groupe. Les estimations qui en résultent se rapprochent passablement de l'estimateur VI original et leur erreur-type est légèrement plus faible, peut-être parce qu'on s'est servi des données complémentaires sur le sexe, l'âge et le degré de scolarité (voir l'analyse, plus loin). Fait intéressant, la désagrégation atténue les effets du rajustement d'une valeur relativement faible dans chaque cas. Il se pourrait qu'en s'écartant de (A4), on ait tendance à trop rajuster l'estimateur VI et que l'approche par désagrégation retenue ici contribue à atténuer le biais et permette d'évaluer la sensibilité des résultats selon la spécification du modèle, quand la désagrégation repose sur d'autres variables.

Ainsi qu'on a pu le lire à la partie 3, les estimations VI se retrouvent souvent à la limite de l'intervalle [0,1]. En réalité, parmi les analyses présentées au tableau 3, seule l'analyse par désagrégation donne des estimations limites. Pour les 64 paramètres $pr(x = i, y = j$, sous-groupe) de $i, j = 1, 2$, sous-groupe $= 1, ..., 16$, cinq estimations se retrouvent à la limite (contre aucune pour les 18 autres paramètres $pr(W = 1 \mid X = 1)$, et le reste). L'erreur-type indiquée au tableau 3 suppose que les paramètres sont connus, si bien qu'on pourrait sous-estimer l'incertitude de l'estimation des paramètres agrégés $pr(x = i, y = j)$.

Le tableau 4 donne d'autres estimations de l'erreur-type pour un sous-groupe, soit les hommes de 26 à 35 ans sans études collégiales. L'estimation de $pr(x = 1, y = 2)$ et celles qui en dérivent, notamment $pr(y = 1 \mid x = 1)$, se situent à la limite. Comme au tableau 3, l'estimation "type" de l'erreur-type repose sur la matrice des observations recueillies et on considère que les paramètres estimés à la limite sont connus. Les estimations bootstrap de l'erreur-type (pour 10,000 répétitions) se rapprochent beaucoup des estimations types pour les paramètres dont l'estimation ne se retrouve pas à la limite. Ainsi on ne possède pas d'estimation type de l'erreur-type pour l'estimation VI de $pr(x = 1, y = 2)$ située à la limite. En fait, estimer l'écart-type de la distribution de l'échantillon aurait peu de sens dans un tel cas. Il serait plus sensé d'établir un intervalle de confiance unilatéral ce qu'on peut faire avec la méthode du profil de vraisemblance, qui donne [0.016], ou avec la méthode bootstrap du percentile qui donne [0.009]. Les intervalles correspondant pour $pr(y = 1 \mid x = 1)$ sont [0.983,1] et [0.990,1].

Tableau 4
Autres estimations de l'erreur-type pour les hommes de 26 à 35 ans sans études collégiales

Paramètre	Estimations VI	Erreur-type estimative: Type	Bootstrap
$pr(W = 1 \mid x = 1)$	0.947	0.011	0.011
$pr(W = 1 \mid x = 2)$	0.107	0.089	0.091
$pr(X = 1 \mid x = 1)$	0.969	0.006	0.007
$pr(X = 1 \mid x = 2)$	0.084	0.088	0.075
$pr(x = 1, y = 1)$	0.953	0.011	0.012
$pr(x = 1, y = 2)$	0	*	*
$pr(x = 2, y = 1)$	0.006	0.007	0.006
$pr(x = 2, y = 2)$	0.041	0.012	0.011
$pr(x = 1)$	0.953	0.011	0.011
$pr(y = 1 \mid x = 1)$	1	*	*
$pr(y = 1 \mid x = 2)$	0.128	0.139	0.117

Nota: $n = 455$; les estimateurs «types» reposent sur la matrice des informations recueillies en vertu de laquelle on présume que les paramètres à la limite sont connus; 10,000 répétitions pour la méthode bootstrap; hypothèses multinomiales.

5. CONCLUSION

L'erreur de mesure peut introduire un biais important dans l'estimation type des taux de transition issus de données longitudinales. Si on peut estimer les taux d'erreur de classification de façon indépendante, il est possible de recourir à diverses méthodes de rajustement. Le présent article montre comment ajuster l'erreur de mesure par l'estimation des variables instrumentales en l'absence de telles données.

Comme c'est le cas pour la méthode d'estimation classique des variables instrumentales, la principale difficulté consiste à trouver une variable qui satisfera les conditions d'une variable instrumentale. En outre, même quand ces conditions sont satisfaites, il est bon que la variable présente un lien robuste avec la situation réelle si on veut obtenir des valeurs raisonnablement précises. Lorsqu'on trouve une variable de

Tableau 2
Taille de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté (échantillonnage multinomial)

Paramètre estimé	Valeur présumée du V de Cramér hypothétique pour les estimateurs VI						
	1.0	0.74	0.59	0.42	0.34	0.24	0.17
	Taille n de l'échantillon requise						
$\mathrm{pr}(x=1, y=1)$	28	59	132	300	320	971	1273
$\mathrm{pr}(x=1, y=2)$	31	50	91	184	219	573	843
$\mathrm{pr}(x=2, y=1)$	1	20	51	129	198	476	811
$\mathrm{pr}(x=2, y=2)$	112	227	366	720	1184	2397	5070
$\mathrm{pr}(y=1 \mid x=1)$	42	60	97	183	219	541	818
$\mathrm{pr}(y=1 \mid x=2)$	57	81	121	216	281	633	1061

4.2 Résultats avec les variables instrumentales réelles

Les résultats de la partie qui précède reposaient sur des variables instrumentales hypothétiques. Pour illustrer de façon plus réaliste notre propos, nous envisagerons maintenant des variables instrumentales réelles éventuelles. La principal difficulté consiste à trouver une variable W qui respecte les hypothèses (A3) et (A4). Il est apparemment plus facile de trouver une variable qui satisfera l'hypothèse (A3) que (A4), essentiellement parce qu'il est plausible que beaucoup de variables quantifiables sans erreur se plient à l'hypothèse (A3). Des variables sans lien avec un changement de la situation d'emploi, donc qui respectent l'hypothèse (A4), paraissent plus difficiles à trouver.

Aux fins d'illustration, nous avons envisagé deux possibilités. Tout d'abord, nous avons retenu la possession d'une voiture pour W (W = 2 si l'intéressé possède une voiture, W = 1 dans le cas contraire). Il se peut que la mesure de cette variable présente une erreur, mais a priori, il est raisonnable de supposer que l'erreur en question n'est pas liée à l'erreur de mesure concernant la situation d'emploi. Ainsi, lorsqu'ils analysent les erreurs résultant de l'inscription de la possession d'une voiture lors du recensement britannique de 1981, Britton et Birch (1985, p. 67) concluent que la principale difficulté associée au faible nombre d'aberrations dérive des véhicules hors d'usage ou disponibles de façon temporaire - ceux loués, par exemple - mais on peut au moins supposer qu'en réalité, ces erreurs présentent peu de liens avec le genre d'erreurs qui survient quand on enregistre la situation d'emploi. Parallèlement, la possession d'une voiture pourrait servir de valeur de remplacement à la situation sociale ou économique associée à un changement de la situation d'emploi, de sorte que l'hypothèse (A4) semble plus douteuse. Dans notre exemple, nous supposons toutefois qu'il y a vérification de (A3) et (A4).

Comme deuxième exemple, nous avons pris pour W la situation d'emploi déphasée en 1985. Dans ce cas, la difficulté est que l'hypothèse (A4) suppose effectivement un processus de Markov pour l'historique de l'emploi et que les taux de transition sont identiques alors qu'en réalité, ils varient d'une personne à l'autre, ce qui infirme l'hypothèse (A4) (lire van de Pol et Langeheine 1990). Pour pouvoir s'écarter de l'hypothèse (A4), il faut désagréger l'échantillon en 16 groupes, définis par classement croisé entre l'âge (quatre groupes), le sexe et le degré de scolarité (études collégiales ou non). On suppose ensuite que le modèle s'applique à l'intérieur d'un sous-groupe et on se sert des tests du rapport de vraisemblance pour déterminer quels paramètres restent constants d'un sous-groupe à l'autre. Ces tests ne donnent qu'une idée très générale, car ils ne tiennent pas compte du plan d'échantillonnage de l'EPDR. On n'a pas relevé de preuve sensible d'écart dans les probabilités d'erreur de classification K_{ij} d'un sous-groupe à l'autre. Il n'existe pas non plus d'indication sensible d'écart dans $\mathrm{Pr}(W \mid x$, sous-groupe) entre les deux sous-groupes du degré de scolarité, pour chacun des huit sous-groupes définis par l'âge et le sexe. En supposant l'égalité des paramètres, on obtient une valeur chi carré non significative du rapport de vraisemblance de la qualité de l'ajustement égale à 52.9 pour 46 dl (46 correspondant au nombre de cellules $= 16\times8 = 128$, moins les deux paramètres K_{ij}, moins $16\times4 = 64$ $\mathrm{pr}(x, y, \text{sous-groupe})$ paramètres moins $8\times2 = 16$ pr $(W \mid x, \text{sous-groupe})$ paramètres). Lorsqu'on combine les paramètres estimatifs du modèle désagrégé, on obtient l'estimation du flux global $\mathrm{pr}(x,y)$.

Le tableau 3 présente l'estimation des paramètres principaux pour les deux sortes de variables instrumentales et la version désagrégée du second choix. On remarque d'abord que l'erreur-type de l'estimateur VI reposant sur la possession d'une voiture est relativement élevée. Le tableau 1 le laissait prévoir, car x et W sont faiblement liés (le V de Cramér est de 0.12). Malgré cela, les rajustements résultants, qui relèvent la valeur estimative des entrées diagonales, sont plausibles et les intervalles de confiance issus de cet estimateur VI semblent plus réalistes que ceux de l'estimateur non rajusté.

L'erreur-type de la deuxième variable instrumentale choisie est plus faible, ainsi qu'on pouvait s'y attendre puisque le lien avec X est plus fort (V de Cramér de 0.73). En

hypothèses (A1) à (A5), est 0.71. On doit la comparer à la valeur hypothétique de 0.75 pour $\mathrm{pr}(x=1, y=1)$. Le biais est donc de $0.71 - 0.75 = -0.04$. Tel qu'indiqué précédemment, on suppose que le biais des estimateurs VI est égal à zéro. L'erreur-type des estimateurs non rajustés vient de la formule binomiale habituelle. Par exemple, l'erreur-type de l'estimateur non rajusté de $\mathrm{pr}(x=1, y=1)$ est $\sqrt{0.71 \times 0.29/5{,}357} = 0.0062$, où 0.71 est la valeur de $\mathrm{Pr}(X=1, Y=1)$. On obtient l'erreur-type des estimateurs VI en prenant la valeur inverse de la matrice d'information prévue, donnée par $n \sum p_{ijk} H_{ijk}$, où H_{ijk} est la matrice 7×7 des dérivées secondes de $\log p_{ijk}$ pour les sept paramètres libres. Après différenciation, on donne aux paramètres leur valeur hypothétique, tel qu'indiqué précédemment. Notons que l'erreur-type obtenue grâce à la matrice d'information multinomiale est sans doute sous-estimée étant donné le plan d'échantillonnage complexe de l'EPDR.

L'erreur-type de l'estimateur VI suit une tendance évidente puisqu'elle augmente à mesure que le lien entre W et x s'affaiblit. La hausse est assez semblable pour tous les paramètres, par exemple le ratio de V = 0.20 contre V = 1.00 se situe entre 3 et 4 pour l'ensemble des paramètres. Dans tous les cas, l'erreur-type de l'estimateur VI est supérieure à celle de l'estimateur non rajusté. La perte d'efficacité entre l'estimateur VI "idéal" (association parfaite entre W et x) et l'estimateur rajusté se situe entre ces deux pôles. Grosso modo, la perte est plus importante pour les paramètres conditionnels que pour les paramètres inconditionnels. Cette perte d'efficacité peut être vue comme la conséquence du rajustement résultant de l'erreur de mesure de y, laquelle reste nécessaire même quand W mesure x à la perfection. En vertu d'une telle interprétation, il est plausible que l'efficacité des paramètres conditionnels diminue davantage, car ils "dépendent moins" des paramètres de la distribution des x marginaux que les données sur W permettent d'estimer.

Pour étudier le compromis entre le biais de l'estimateur non rajusté et la plus forte variance de l'estimateur VI, nous avons calculé la taille minimale n de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté. Pour les plans d'échantillonnage complexes, on suppose que la taille de l'échantillon correspond à la taille de l'échantillon après avoir prix en compte l'effet de plan. Le tableau 2 indique la valeur minimale de l'échantillon pour des forces d'association variables entre W et x. Sans erreur de classification, toutes les entrées se trouvent à l'infini puisque l'efficacité des estimateurs non rajustés est toujours supérieure à celle des estimateurs VI. Pour un nombre hypothétique d'erreurs, délimité par $K_{21} = 0.03$ et $K_{12} = 0.06$, la taille de l'échantillon requise augmente rapidement à mesure que V diminue. La variation dans les rangées du tableau 1 et la variation du biais des estimateurs non rajustés expliquent en partie l'écart entre les rangées du tableau 2. Le biais de l'estimateur non rajusté de $\mathrm{pr}(x=2, y=2)$ est donc relativement faible, ce qui se traduit par les fortes valeurs de la rangée correspondante, au tableau 2. Soulignons que la valeur de 1 pour $\mathrm{pr}(x=2, y=1)$ et pour le facteur V de Cramér résulte du fait que les deux estimateurs ont la même erreur-type (voir le tableau 1), si bien que le biais des estimateurs non rajustés suppose que l'EQM de l'estimateur VI est plus faible pour $n \geq 1$.

La principale conclusion qu'on retire du tableau 2 reste que dans un certain nombre de situations pratiques, l'estimation des VI s'avère utile pourvu que les hypothèses du modèle soient valables, même s'il faut légèrement augmenter la taille de l'échantillon afin d'accepter des plans d'échantillonnage complexes.

Tableau 1
Biais et erreur-type sous diverses VI hypothétiques

			Valeur des paramètres hypothétiques de l'estimateur VI						
$\mathrm{pr}(W=1 \mid x=1)$			1.0	0.1	0.1	0.1	0.3	0.1	0.5
$\mathrm{pr}(W=1 \mid x=2)$			0.0	0.9	0.7	0.5	0.7	0.3	0.3
V de Cramér			1.0	0.74	0.59	0.42	0.34	0.24	0.17
		Valeur des paramètres hypothétiques de l'estimateur VI							
Paramètre estimatif	Bias (× 100) de l'estimateur non rajusté	Estimateur non rajusté	Estimateur VI						
$\mathrm{pr}(x=1, y=1)$	−4.0	0.62	0.68	0.75	0.88	1.13	1.16	1.82	2.05
$\mathrm{pr}(x=1, y=2)$	3.0	0.32	0.39	0.43	0.51	0.64	0.69	1.03	1.24
$\mathrm{pr}(x=2, y=1)$	3.0	0.32	0.32	0.37	0.44	0.57	0.66	0.95	1.27
$\mathrm{pr}(x=2, y=2)$	−2.0	0.51	0.59	0.65	0.73	0.89	1.06	1.42	1.99
$\mathrm{pr}(y=1 \mid x=1)$	−3.9	0.37	0.55	0.55	0.64	0.81	0.88	1.30	1.58
$\mathrm{pr}(y=1 \mid x=2)$	12.4	0.60	1.40	1.63	1.95	2.56	2.90	4.30	5.55

Nota: 1 = occupé, 2 = non occupé, $n = 5{,}357$; échantillonnage multinomial; biais des estimateurs VI égal à zéro.

Dans le reste de l'article, nous n'examinerons que le cas $r = s = 2$ en vertu du quel le modèle vient d'être identifié (sauf pour les valeurs aberrantes des paramètres). On pourrait établir que $p_{ijk} = n_{ijk}/n$ et résoudre les équations (6) et (7) pour les paramètres inconnus. Si elles se retrouvent dans les limites autorisées, c'est-à-dire dans la fourchette [0,1], les solutions résultantes correspondront aux estimations des VI. Dans la pratique cependant, on se rend compte que les échantillons de taille moyenne débouchent sur des solutions souvent irréalisables. De plus, les calculs sont complexes. Nous avons jugé plus facile de maximiser l directement au moyen des méthodes numériques de l'ensemble GAUSS (Edlefsen et Jones 1984) ou d'ensembles qui conviennent aux modèles à structure latente reposant sur l'algorithme EM, tel PANMARK (van de Pol, Langeheine et de Jong 1991). Un ensemble à structure latente permettrait d'ajuster un modèle à deux classes sans restriction, puis d'estimer θ_1 et θ_2 avec (7). Rien ne garantira néanmoins que les estimations résultantes tombent dans la fourchette [0,1] acceptable. S'ajouterait en outre la complication qu'on doive établir l'erreur-type de l'estimation de θ_1 et de θ_2 à partir de la matrice des covariances des estimations de $(R_1, R_2, K_{21}, K_{22})$. Il nous a semblé plus commode d'ajuster directement le modèle comme un modèle restreint à structure latente. Un autre avantage de cette approche est qu'elle s'étend naturellement à l'ajustement de modèles analogues d'un sous-groupe à l'autre, sous réserve de la contrainte éventuelle que certains paramètres restent constants entre les sous-groupes. Nous y reviendrons à la partie 4.

Avec les hypothèses multinomiales, il se pourrait que l'erreur-type repose sur la dérivée seconde du logarithme du rapport de vraisemblance obtenu pour l'estimation des VI. Pareille approche devient toutefois problématique quand la valeur maximale de l se trouve à la limite de l'espace paramétrique. Une solution consiste simplement à prendre les valeurs des paramètres telles que connues, à la limite. Le risque est qu'on sous-estime l'incertitude. Baker et Laird (1988) examinent deux autres approches au calcul de l'intervalle estimatif des paramètres dans des circonstances de ce genre: une méthode bootstrap et la méthode du profil de vraisemblance. La première méthode suppose le prélèvement réitéré d'échantillons multinomiaux où p_{ijk} est égal à n_{ijk}/n, et l'inscription de la distribution des paramètres estimatifs pour une suite d'échantillons bootstrap. La méthode du profil de vraisemblance donne les intervalles estimatifs d'un paramètre sous la forme des jeux de valeurs du paramètre non rejetées par un test du ratio de vraisemblance. Ces méthodes sont illustrées à la fin de la partie 4.

4. EXEMPLES NUMÉRIQUES

Aux fins d'illustration numérique, nous nous servirons des données du sous-échantillon à probabilité identique de l'étude américaine par panel sur la dynamique du revenu (EPDR) (lire Hill 1992). Nous examinerons deux états, occupé et non occupé, codés 1 et 2 respectivement, donc limiterons encore une fois notre attention au cas binaire. Pour plus de commodité, nous ne tiendrons pas compte des non-réponses et prendrons l'échantillon de 5,357 personnes de 18 à 64 ans de 1986 avec les valeurs entières pour les variables situation d'emploi en 1985, en 1986 et en 1987, possession d'une voiture, âge, sexe et degré de scolarité.

Les propriétés de l'estimateur VI sont évaluées de deux façons. Tout d'abord, à la partie 4.1, on compare le biais et l'erreur-type de l'estimateur VI à ceux de l'estimateur "non rajusté" des variables instrumentales hypothétiques, pour diverses associations avec x. Deuxièmement, à la partie 4.2, on étudie les conséquences de l'utilisation de différentes variables réelles de l'EPDR comme variables instrumentales.

4.1 Propriétés du biais et de l'erreur-type des estimateurs des variables instrumentales hypothétiques

Les paramètres qui nous intéressent le plus sont les probabilités conjointes $\mathrm{pr}(x = i, y = j)$ ou les probabilités conditionnelles $\mathrm{pr}(y = j \mid x = i)$ qui en dérivent. Les estimateurs simples "non rajustés" des paramètres reposent sur les proportions de l'échantillon correspondantes des variables X et Y classées et ont pour espérance $\mathrm{pr}(X = i, Y = j)$, pour un échantillonnage multinomial. Puisque $\mathrm{Pr}(X = i, Y = j)$ diffère généralement de $\mathrm{pr}(x = i, y = j)$, les estimateurs non rajustés sont habituellement biaisés. En supposant que les hypothèses (A1) à (A5) du modèle se vérifient, les estimateurs VI de $\mathrm{pr}(x = i, y = j)$ ne seront pas biaisés de façon asymptotique, même avec une variance supérieure à celle des estimateurs non rajustés. Nous essayerons ici de déterminer l'importance du compromis entre le biais des estimateurs non rajustés et le relèvement de la variance des estimateurs VI. On supposera que les hypothèses du modèle (A1) à (A5) tiennent et que l'échantillon est assez important pour que l'estimateur VI ne présente pas de biais.

Dans le cadre de notre analyse numérique, il est préférable d'utiliser certaines valeurs "réalistes" des paramètres. Celles-ci viennent de l'arrondissement des estimations du flux annuel survenu entre 1986 et 1987 selon les analyses de la partie 4.2 (voir le tableau 3). Les valeurs des cinq paramètres libres du modèle indépendant de W qui ont été retenues sont $K_{21} = 0.03$, $K_{22} = 0.94$, $\mathrm{pr}(x = 2) = \pi = 0.22$, $\mathrm{pr}(y = 2, x = 1) = \theta_1(1 - \pi) = 0.03$ et $\mathrm{pr}(y = 2, x = 2) = \theta_2 \pi = 0.19$. Les valeurs des deux autres paramètres libres $\varphi_{11} = \mathrm{pr}(W = 1 \mid x = 1)$ et $\varphi_{12} = \mathrm{pr}(W = 1 \mid x = 2)$ apparaissent dans des colonnes différentes du tableau 1. La statistique V de Cramér, qui définit le lien entre deux variables binaires en répartissant essentiellement la valeur chi carré sur un intervalle [0,1], résume la robustesse du lien entre les variables W et x. Pour chaque valeur possible des paramètres, le tableau 1 donne l'erreur-type estimative des estimateurs VI pour un échantillon $n = 5{,}357$ de l'EPDR. Le tableau 1 indique aussi le biais et l'erreur-type de l'estimateur non rajusté pour les mêmes valeurs paramétriques $K_{21}, K_{22}, \pi, \theta_1$ et θ_2, avec un échantillon de taille identique.

Pour montrer comment on calcule le biais des estimateurs non rajustés, considérons $\mathrm{pr}(x = 1, y = 1)$. L'espérance de l'estimateur non rajusté du paramètre, soit $\mathrm{pr}(X = 1, Y = 1)$, qu'on obtient à partir des valeurs $K_{21}, K_{22}, \pi, \theta_1$ et θ_2 et des

$2(r-1)r^2+(s-1)r^2+(r^2-1)$ paramètres donnés par les $(r-1)r^2$ paramètres $\mathrm{pr}(X=i \mid x=u, y=v)$, les $(r-1)r^2$ paramètres $\mathrm{pr}(Y=j \mid x=u, y=v)$, les $(s-1)r^2$ paramètres $\mathrm{pr}(W=k \mid x=u, y=v)$ et les r^2-1 paramètres libres $\mathrm{pr}(x=u, y=v)$. Ces paramètres sont soumis aux $2r(r-1)^2$ restrictions de (A2) et aux $(s-1)r(r-1)$ restrictions que suppose (A4). Nous nous attaquerons d'abord au cas $r=2$. Dans ce cas, $4s+7$ paramètres sont sujets à $2s+2$ restrictions, ce qui laisse $2s+5$ paramètres libres

$$\left\{K_{x2u}, K_{y2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u=1,2\ v=1,2\right\},$$

où $\varphi_{ku}=\mathrm{pr}(W=k \mid x=u)$, $\theta_u=\mathrm{pr}(y=2 \mid x=u)$, et $\pi=\mathrm{pr}(x=2)$. Le nombre probable de cellules «libres» dans la grille de X par Y par W est r^2s-1, ou $4s-1$ quand $r=2$. Une condition essentielle à l'identification quand $r=2$ est donc que $4s-1 \geq 2s+5$ ou $s \geq 3$. Malheureusement, cette condition ne suffit pas. Soit

$$R_u=\mathrm{pr}(Y=2 \mid x=u)=\sum_{v=1}^{2} K_{y2v}\,\theta_u^{v-1}(1-\theta_u)^{2-v}, \qquad (3)$$

alors

$$\mathrm{pr}(X=i, Y=j, W=k)=$$
$$\sum_{u=1}^{2} K_{xiu}\varphi_{ku}R_u^{j-1}(1-R_u)^{2-j}\pi^{u-1}(1-\pi)^{2-u}. \qquad (4)$$

Les $4s-1$ probabilités des cellules libres sont donc déterminées par seulement les $2s+3$ paramètres

$$\left\{K_{x2u}, \varphi_{2u}, \ldots, \varphi_{su}, R_u, \pi;\ u=1,2\right\}$$

qui ne seront identifiés que si $4s-1 \geq 2s+3$ ou $s \geq 2$. En réalité, cette condition suffit pour identifier les paramètres, sauf dans quelques combinaisons exceptionnelles. (Lire Madansky (1960) pour le cas où $s=2$, et Goodman (1974) pour le cas général où $s \geq 2$.)

Toutefois, bien que les $2s+3$ paramètres qui précèdent soient généralement identifiés lorsque $s \geq 2$, on ne peut déterminer les quatre paramètres K_{y21}, K_{y22}, θ_1 et θ_2, car ils ne sont associés qu'à deux paramètres connus, R_1 et R_2, au moyen de l'équation (3). Les paramètres principaux qui nous intéressent, notamment θ_1 et θ_2, restent mal identifiés, peu importe la valeur de s.

Pour les raisons qui précèdent, il faut imposer au moins deux autres restrictions au modèle si l'on veut identifier θ_1 et θ_2. À l'instar de Chua et Fuller (1987), on pourrait supposer que les erreurs de mesure ne sont pas biaisées comme dans (2), ce qui impose deux contraintes:

$$\pi=K_{x21}(1-\pi)+K_{x22}\pi \qquad (5)$$

$$\theta_1(1-\pi)+\theta_2\pi=R_1(1-\pi)+R_2\pi. \qquad (6)$$

Malheureusement, la première contrainte ne s'applique qu'aux paramètres déjà identifiés lorsque $s \geq 2$, si bien qu'on ne peut identifier θ_1 et θ_2. Nous formulerons donc une hypothèse supplémentaire, c'est-à-dire que le processus d'erreur est constant dans le temps, de sorte que

$$K_{xiu}=K_{yiu}=K_{iu}, \text{ par exemple, pour } i,u=1,2,\ldots,r. \qquad (A5)$$

Pareille hypothèse semble élémentaire et naturelle si on applique la même procédure de mesure dans le cadre de l'enquête d'année en année. L'hypothèse élimine le problème de sous-identification du cas $r=2$ examiné précédemment, puisqu'en identifiant $K_{xiu}=K_{iu}$ et R_u, on peut déterminer θ_u à partir de (3)

$$\theta_u=(R_u-K_{21})/(K_{22}-K_{21}) \qquad (7)$$

(en excluant le cas ordinaire où les variables quantifiées sont indépendantes des variables réelles, si bien que $K_{22}=K_{21}$).

En résumé, quand les hypothèses (A1) à (A5) se confirment et que $r=2$, le modèle comporte $2s+3$ paramètres libres $\{K_{2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u=1,2\}$ que l'on peut identifier si $s \geq 2$, sauf dans les cas exceptionnels comme ceux analysés par Madansky (1960).

Enfin, revenons au cas général r. Puisque (A5) impose $(r-1)r$ restrictions, il s'ensuit qu'on compte $2(r-1)r^2+(s-1)r^2+(r^2-1)-[2r(r-1)^2+(s-1)r(r-1)]-(r-1)r=2r^2+sr-2r-1$ paramètres libres. Il y a donc r^2s-1 cellules libres probables dans le tableau de X par Y par W et, dans l'ensemble, on pourra identifier le modèle si $r(r-1)(s-2) \geq 0$. La condition relative à l'identification des paramètres reste donc $s \geq 2$, pour une valeur quelconque de $r \geq 2$. On peut dire d'autre part que

$$R_{ju}=\mathrm{Pr}(Y=j \mid x=u)=\sum_{v=1}^{r} K_{jv}\theta_{uv}$$

où $\theta_{uv}=\mathrm{pr}(y=v \mid x=u)$. Par conséquent, il est possible de déterminer θ_{uv} à partir de R_{ju} et K_{jv}, également identifiés, pourvu que la matrice $[K_{iu}]$ ne soit pas singulière. Pour le cas général r, le modèle est identifié en vertu des hypothèses (A1) à (A5), sauf dans les cas exceptionnels examinés par Goodman (1974).

3. ESTIMATION

Supposons que pour un échantillon de taille n on retrouve le chiffre n_{ijk} dans les cellules du tableau de contingence $r \times r \times s$ de $X \times Y \times W$, et que les cellules aient une distribution multinomiale selon les paramètres n et $p_{ijk}=\mathrm{pr}(X=i, Y=j, W=k)$. La vraisemblance logarithmique implicite est

$$l=\sum_i \sum_j \sum_k n_{ijk}\,\log p_{ijk}.$$

En vertu d'un plan d'échantillonnage complexe, on peut considérer que n_{ijk} est un chiffre pondéré, ce qui débouche sur une pseudo-vraisemblance logarithmique (Skinner 1989). Les estimateurs des paramètres obtenus lorsqu'on maximise l porteront le nom d'estimateurs de *variables instrumentales* (VI).

Pareille hypothèse se retrouve couramment dans les modèles généraux à variables latentes (lire Anderson 1959). Elle semble raisonnable quand les méthodes de mesure de l'enquête sont indépendantes aux deux occasions. Si on obtient X rétrospectivement de la même interview qui a servi à mesurer Y, il est raisonnable de croire que la tendance des répondants à fournir des réponses exagérément cohérentes lors d'une interview entraînera une association positive entre les erreurs de classification. Marquis et Moore (1990) le prouvent avec l'enquête sur le revenu et la participation aux programmes. La possibilité d'une hétérogénéité individuelle dans les probabilités d'erreur de classification (certains répondants pouvant se montrer plus fiables que d'autres) est une deuxième raison pour questionner l'hypothèse d'indépendance conditionnelle. On lira à ce sujet Skinner et Torelli (1993), et Singh et Rao (1995). À la partie 4, nous accepterons l'hétérogénéité en supposant seulement que le modèle se confirme dans les cellules d'un tableau des variables observées.

La seconde hypothèse fondamentale est que l'erreur de classification ne dépend que de l'état réel courant, si bien que

$$\mathrm{pr}(X = i \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u) = K_{xiu}, \quad \text{(A2)}$$
par exemple, et
$$\mathrm{pr}(Y = j \mid x = u, y = v) = \mathrm{pr}(Y = j \mid y = v) = K_{yjv}, \text{ par exemple.}$$

K_{xiu} et K_{yjv} définissent les matrices $r \times r$ des erreurs de classification $\boldsymbol{K}_x = [K_{xiu}]$ et $\boldsymbol{K}_y = [K_{yjv}]$. Soit $\boldsymbol{P}$, la matrice $r \times r$ ayant pour ij-ième élément $\mathrm{pr}(X = i, Y = j)$ et Π, la matrice $r \times r$ dont le uv-ième élément est $\mathrm{pr}(x = u, y = v)$. On obtient l'équation matricielle

$$\boldsymbol{P} = \boldsymbol{K}_x \Pi \boldsymbol{K}_y'. \quad (1)$$

La matrice Π renferme les paramètres auxquels on s'intéresse alors que c'est la matrice $\boldsymbol{P}$ qu'on peut estimer à partir des valeurs X et Y venant de l'échantillon. Si on dispose des estimations complémentaires $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ et qu'il ne s'agit pas de matrices singulières, il est possible de résoudre l'équation (1) pour estimer Π. Par ailleurs, si on peut vérifier l'état réel grâce à une réinterview, on peut estimer directement $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ (Abowd et Zellner 1985). D'un autre côté, si on n'obtient qu'une nouvelle classification indépendante à la réinterview, on ne pourra estimer que les matrices interview-réinterview

$$\boldsymbol{K}_x \Delta_x \boldsymbol{K}_x' \text{ et } \boldsymbol{K}_y \Delta_y \boldsymbol{K}_y'$$

où $\Delta_x = \mathrm{diag}[\mathrm{pr}(x = u)]$ et $\Delta_y = \mathrm{diag}[\mathrm{pr}(y = v)]$ (Chua et Fuller 1987). Chaque matrice interview-réinterview est symétrique et la somme de leurs éléments donne un. Elles ne renferment donc que $r(r+1)/2 - 1$ éléments d'information «indépendants». Puisque chaque colonne de la matrice $\boldsymbol{K}$ et la diagonale de la matrice Δ ont pour somme un, le nombre de paramètres inconnus à chaque occasion est $r(r-1) + r - 1 = r^2 - 1$. Par conséquent, la différence entre le nombre de paramètres et le nombre d'éléments d'information est $r^2 - 1 - r(r+1)/2 + 1 = r(r-1)/2$ à chaque occasion, et le modèle est mal identifié lorsque $r \geq 2$. Chua et Fuller (1987) proposent qu'on ajoute une hypothèse naturelle afin de faciliter l'identification, à savoir supposer que les erreurs de mesure ne sont pas biaisées ou que

$$\mathrm{pr}(x = i) = \mathrm{pr}(X = i), \mathrm{pr}(y = i) = \mathrm{pr}(Y = i)\, i = 1, ..., r. \quad (2)$$

Dans ce cas, les valeurs faussement positives et faussement négatives ont tendance à s'annuler lors de l'estimation transversale des proportions. La nouvelle hypothèse réduit le nombre de paramètres de $r - 1$ à chaque occasion. Malgré cela, le modèle reste mal identifié lorsque $r \geq 3$, et Chua et Fuller (1987) sont forcés d'introduire de nouvelles hypothèses.

Voyons maintenant comment identifier le modèle sans les données d'une réinterview. Dans une régression linéaire simple où la covariable comporte une erreur de mesure la covariable, l'approche de la variable instrumentale (Fuller 1987, partie 1.4) suppose l'existence d'une variable instrumentale «observée» W, corrélée à la covariable mais indépendante de l'erreur de mesure et de l'erreur de l'équation de régression. Nous avons élargi cette hypothèse au cadre actuel en faisant en sorte que W soit une *variable instrumentale* si elle n'est pas indépendante de x et si

W et (X, Y) sont conditionnellement indépendants étant donné (x,y), (A3)
si W et y sont conditionnellement indépendants étant donné x. (A4)

En général, nous laisserons W être une variable nominale comprenant un nombre arbitraire s de catégories même si, en pratique, $s = r$ puisqu'on désire que W soit étroitement lié à x. Une possibilité spécifique consiste à faire correspondre W à l'état classé au temps $t - 1$. L'usage de la valeur déphasée d'une «covariable» comme variable instrumentale remonte aux débuts des travaux sur l'estimation des variables instrumentales (lire Reiersol 1941; Durbin 1954). Dans le cas présent, l'hypothèse A4 se confirme si les états réels suivent un processus de Markov et si les erreurs de classification sont conditionnellement indépendantes comme dans A1.

On peut représenter le modèle issu des hypothèses (A1) à (A4) par le graphique d'indépendance conditionnelle de la figure 1. Chaque point représente une variable. Les points ne sont pas reliés si les variables correspondantes présentent une indépendance conditionnelle avec les autres variables.

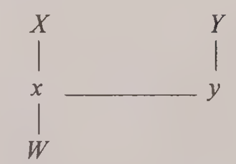


Figure 1. Représentation graphique de l'indépendance conditionnelle dans le modèle de base.

Le modèle illustre un modèle restreint à structure latente (Goodman 1974) où les variables X, Y et W observées sont conditionnellement indépendantes étant donné les variables latentes x et y, bref sont indépendantes dans les r^2 classes latentes définies par les couples (x,y). Ce modèle comprend

Techniques d'enquête, Juin 1997
Vol. 23, n° 1, pp. 61-68
Statistique Canada

Estimation des variables instrumentales des flux bruts en présence d'erreur de mesure

K. HUMPHREYS et C. J. SKINNER[1]

RÉSUMÉ

Les auteurs étudient le problème qui consiste à estimer les taux de transition au moyen des données d'une enquête longitudinale, lorsqu'il existe des erreurs de classification. Ils examinent les approches faisant appel à des données auxiliaires sur les taux de classification erronée et ainsi que d'autres approches pour modéliser l'erreur de mesure. À partir de variables instrumentales nominales, ils suggèrent comment identifier et estimer les modèles qui comprennent des variables de ce genre en considérant un modèle à structure latente restreinte. Enfin, ils étudient les propriétés numériques des estimateurs implicites des variables instrumentales pour les taux des flux grâce aux données de l'étude par panel sur la dynamique de revenu.

MOTS CLÉS: Structure latente; longitudinal; erreur de classification; taux de transition.

1. INTRODUCTION

Un des principaux avantages d'une enquête longitudinale est qu'elle permet d'estimer les flux bruts, par exemple le nombre de chômeurs qui trouvent un emploi (lire Hogue et Flaim 1986). Quand on estime ces flux, le biais introduit par l'erreur de mesure pose un problème majeur. En effet, si les erreurs de classification peuvent avoir tendance à s'annuler lorsqu'on estime des proportions transversales (Chua et Fuller 1987), on ne peut en dire autant quand il s'agit de flux longitudinaux.

La première tentative en vue de résoudre le problème de l'erreur de mesure devrait clairement être de réduire cette erreur dans les méthodes de mesure de l'enquête. Biemer, Groves, Lyberg, Mathiowetz et Sudman (1991) examinent les approches pertinentes, aussi ne les reprendrons-nous pas ici. Néanmoins, aussi bonnes que puissent être les méthodes d'enquête, l'erreur de mesure reste inévitable et on devra en compenser les effets à l'analyse.

En général, les méthodes utilisées pour compenser l'erreur de mesure reposent sur un modèle hypothétique quelconque du processus d'erreur. Plusieurs auteurs décrivent divers modèles qu'on examinera à la partie 2. Pour les identifier et les estimer, on a habituellement besoin d'informations complémentaires, par exemple les renseignements recueillis lors des réinterviews (lire Meyer 1988). Les réinterviews coûtant cher et leur but étant rarement d'estimer les paramètres de la distribution des erreurs de mesure dans la pratique (Forsman et Schreiner 1991), d'autres méthodes s'avèrent nécessaires en l'absence de données issues des réinterviews. Quand l'erreur de mesure porte sur des variables continues et qu'on manque d'informations auxiliaires sur la distribution de cette erreur, on recourt couramment à l'estimation des variables instrumentales (lire Fuller 1987, partie 1.4). Par variable instrumentale, on entend une variable faisant partie des données d'enquête associées à la variable réelle, comprenant une erreur de mesure, mais non corrélée avec cette dernière. Les variables instrumentales et les hypothèses qui s'y rattachent nous fournissent l'information susceptible de remplacer celle des réinterviews et de faciliter l'identification et l'estimation des paramètres du modèle, parmi lesquels se retrouve la variable réelle. Dans cet article, nous verrons comment adapter la méthode d'estimation des variables instrumentales pour estimer les flux entre états discrets. On se rend compte que les modèles à structure latente (lire Bartholomew 1987, Ch. 2) procurent un cadre général dans lequel les hypothèses sur la variable instrumentale correspondent à certaines restrictions imposées aux paramètres du modèle. Notre approche rappelle donc celles qui imposent des restrictions aux modèles à structure latente (van de Pol et de Leeuw 1986; van de Pol et Langeheine 1990).

2. MODÈLES

Nous n'étudierons que le cas à deux possibilités, $t = 1$ et $t = 2$. Soit r, le nombre d'états dans lequel une personne peut être classée à chaque occasion. Appelons les états classés à $t = 1$ et $t = 2$ X et Y, respectivement, et les états véritables, x et y. On suppose l'existence d'un modèle où la valeur vectorielle de (X, Y, x, y) est le résultat indépendant d'un vecteur aléatoire commun de distribution $\mathrm{pr}(X = i, Y = j, x = u, y = v)$.

La première hypothèse sur cette distribution, formulée par divers auteurs (lire Abowd et Zellner 1985; Poterba et Summers 1986 et Chua et Fuller 1987), et que nous reprendrons ici, est que les erreurs de classification effectuées aux deux occasions sont conditionnellement indépendantes, l'état véritable étant connu, c'est-à-dire

$$\mathrm{pr}(X = i, Y = j \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u, y = v)\,\mathrm{pr}(Y = i \mid x = u, y = v). \quad \text{(A1)}$$

[1] K. Humphreys, Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden; C. J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

rédacteur, pour leurs commentaires précieux qui ont permis d'améliorer considérablement cet article.

BIBLIOGRAPHIE

BABU, G.J. (1986). A note on bootstrapping the variance of sample quantile. *Annals of the Institute of Statistical Mathematics*, 38-A, 439-443.

BEACH, C.M., et DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BEACH, C.M., et KALISKI, S.F. (1986). Lorenz curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics*, 35, 38-45.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., et KOVAČEVIĆ, M.S. (1995). Estimation de l'inégalité du revenu d'après les données d'enquête: application de la méthode des équations d'estimation. *Techniques d'enquête*, 21, 151-159.

BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

EFRON, B. (1979). Bootstrap method: another look at the jackknife. *Annals of Statistics* 7, 1-26.

FOSTER, J.E., et WOLFSON, M.C. (1992). Polarization and the decline of the middle class: Canada and the U.S. (Manuscrit).

FRANCISCO, C.A., et FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KAKWANI, N.C. (1980). *Income Inequality and Poverty*. Washington, D.C.: World Bank.

KOVAČEVIĆ, M.S., YUNG, W., et PANDHER, G.S. (1995). Estimating the Sampling Variances of Income Inequality and Polarization – An Empirical Study. Direction de la méthodologie, Document de travail, HSMD-95-007-E. Statistique Canada.

KOVAČEVIĆ, M.S., et BINDER, D.A. (1997). Variance estimation for measures of income inequality and polarization- the estimating equations approach. (A paraître dans *Journal of Official Statistics*).

KOVAR, J.G. (1987). Variance Estimation of Medians in Stratified Samples. Direction de la méthodologie, document de travail, BSMD-87-004-E. Statistique Canada.

KOVAR, J.G., RAO, J.N.K., et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics* 16, 25-45.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

LOVE, R., et WOLFSON, M.C. (1976). Income inequality: statistical methodology and Canadian illustrations. Ottawa, Statistique Canada.

McCARTHY, P.J. (1993). Standard error and confidence interval estimation for the median. *Journal of Official Statistics*, 9, 673-689.

NYGÄRD, F., et SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist & Wiksell International.

NYGÄRD, F., et SANDSTRÖM, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics*, 1, 399-412.

RAO, J.N.K., et SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

RAO, J.N.K., et WU, C.F.J. (1987). Methods for standard errors and confidence intervals from survey data: Some recent work. *Proceedings of the 46th Session of International Statistical Institute*, 3, 5-19.

RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

SANDSTRÖM, A., WRETMAN, J.H., et WALDÉN, B. (1985). Variance estimators of the Gini coefficient, simple random sampling. *Metron*, 43, 41-70.

SANDSTRÖM, A., WRETMAN, J.H., et WALDÉN, B. (1988). Variance estimators of the Gini coefficient – probability sampling. *Journal of Business and Economic Statistics*, 6, 113-119.

SEN, A.K. (1973). *On Economic Inequality*. London: Oxford University Press.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 119-122.

SHAO, J. (1993). Balanced repeated replication. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 544-549.

SHAO, J., et WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

WOLFSON, M.C. (1994). When inequalities diverge. *American Economic Review*, 84, 353-358.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.

YITZHATI, S. (1991). Calculating jackknife variance estimators for parameters of the Gini method. *Journal of Business and Economic Statistics*, 9, 235-239.

Tableau 5
Classements des méthodes selon le biais relatif, la stabilité relative et la probabilité empirique de couverture

	Jackknife	GDÉC	GRDÉC	Bootstrap	ÉE (Taylor)	Meilleures méthodes
Indice de Gini	Toutes les méthodes donnent des résultats comparables					–
Quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	ÉE, BS
Ordonnées de la courbe	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 3	2, 1, 2	ÉE, BS
Parts de quantiles	5, 5, 5	3, 4, 4	4, 3, 1	1, 2, 2	2, 1, 3	BS, ÉE
Proportion de faible	5, 5, 5	3, 4, 2	4, 3, 1	2, 2 ,3	1, 1, 4	ÉE, BS
Indice de polarisation	5, 5, 5	3, 4, 4	4, 3, 2	2, 1, 1	1, 2, 3	BS, ÉE

Les estimateurs jackknife sont les moins stables. La stabilité des estimateurs GRDÉC, bootstrap et ÉE est comparable et, quand on fait la moyenne pour l'ensemble des quantiles, environ trois fois plus grande que celle des estimateurs jackknife. C'est autour de la médiane que nous avons obtenu la stabilité la plus forte (voir le figure 3b).

En général, la probabilité de couverture calculée pour les quantiles est inférieure à la valeur nominale pour toutes les méthodes étudiées, à quelques exceptions près pour les méthodes GDÉC et GRDÉC (voir figure 3c). En revanche, la comparaison des taux d'erreur des intervalles latéraux semble indiquer que toutes les méthodes donnent des résultats semblables, le taux calculé pour l'intervalle latéral supérieur (droit) étant plus élevé pour les quantiles inférieurs ($p = 0.1, 0.2$); pour d'autres, c'est la situation inverse que l'on observe, la queue inférieure (gauche) étant plus importante. Les résultats obtenus pour les intervalles de confiance de 90% et de 99% sont similaires.

Les résultats de cette étude empirique confirment qu'il faut éviter d'utiliser la méthode jackknife pour estimer la variance des quantiles. Dans le cas de la variance de la médiane en particulier, le choix le meilleur semble être celui de la méthode des équations d'estimation ou de la méthode bootstrap. Pour d'autres quantiles, la méthode du GRDÉC donne également de très bons résultats.

Nos résultats sont condensés dans le tableau 5 où nous classons le biais relatif, la variation relative et la probabilité de couverture obtenus pour les méthodes étudiées selon une échelle allant de 1 à 5 (1 = résultat le meilleur). Pour les méthodes de rééchantillonnage, nous avons calculé la moyenne des valeurs obtenues pour les deux estimateurs. Pour les quantiles, l'ordonnée de la courbe de Lorenz et la part du quantile, nous avons calculé la moyenne des valeurs obtenues pour l'ensemble des p. Dans la dernière colonne, nous indiquons les deux meilleures méthodes.

5. DISCUSSION ET CONCLUSION

La méthode de linéarisation au moyen des équations d'estimation est celle qui donne le meilleur résultat dans l'ensemble, produisant le biais relatif le plus petit, la variation relative la plus faible et d'assez bonnes propriétés de couverture. Vient ensuite la méthode bootstrap, qui est la meilleure des méthodes de rééchantillonnage envisagées. Les méthodes GRDÉC et GDÉC donnent des résultats aussi bons pour l'ordonnée de la courbe de Lorenz, la part du quantile et certains quantiles, en ce sens qu'elles produisent des estimations dont le faible biais relatif et la stabilité relative sont du même ordre que dans le cas de la méthode bootstrap. La méthode jackknife donne des résultats médiocres pour toutes les mesures de l'inégalité du revenu, sauf l'indice de Gini.

Il est bien connu que l'estimation de la variance selon la méthode jackknife donne de mauvais résultats dans le cas de fonctions non lissées. Le degré de lissage de la fonction J définie en (3.1) est un déterminant essentiel des propriétés asymptotiques de l'estimateur de sa variance. Si nous classons les mesures de l'inégalité du revenu dans la catégorie «lisses» ou «non lisses» d'après les fonctions J, nous constatons que le seul estimateur lisse étudié ici est l'indice de Gini. Il n'est donc pas surprenant que ce soit la seule mesure du revenu pour laquelle la méthode jackknife donne de bons résultats. Néanmoins, avant de décider d'utiliser l'estimateur jackknife de la variance, il convient de vérifier que les hypothèses en vertu desquelles la méthode est valide sont satisfaites.

Si l'objectif est de fournir une méthode d'estimation de la variance pour la longue liste de statistiques du revenu, notre étude empirique montre que la méthode bootstrap est la méthode par échantillonnage la plus satisfaisante et que la linéarisation au moyen d'équations d'estimation est non seulement la méthode la meilleure, mais aussi celle demandant le moins de calculs, même si elle nécessite certains travaux algébriques préparatoires, qui diffèrent pour chaque mesure du revenu.

Il convient de souligner que l'étude empirique se fonde sur un plan d'échantillonnage par grappes à un degré, les grappes étant sélectionnées avec probabilité proportionnelle à la taille, de sorte qu'il n'est pas tenu compte de la variabilité à l'intérieur de la grappe. Certaines études de faible portée aboutissent à des observations similaires au sujet des méthodes étudiées ici dans le cas de plans d'échantillonnage à deux degrés (voir Binder et Kovačević, 1995, et Kovačević et Binder, 1997

REMERCIEMENTS

Les auteurs remercient G.S. Pandher pour sa participation fructueuse au démarrage du projet, J. Gambino pour sa lecture minutieuse d'une version antérieure de l'article, ainsi que H. Mantel, rédacteur adjoint, les arbitres anonymes et le

de la courbe de Lorenz, le biais relatif varie de −2% à +3% dans le cas de la méthode bootstrap et de −5% à +1% dans le cas de la méthode des équations d'estimation. Pour la part du quantile, le biais relatif varie de −3% à +8% pour les estimations obtenues par la méthode bootstrap, et de −3% à +5% pour celles obtenues par la méthode des équations d'estimation.

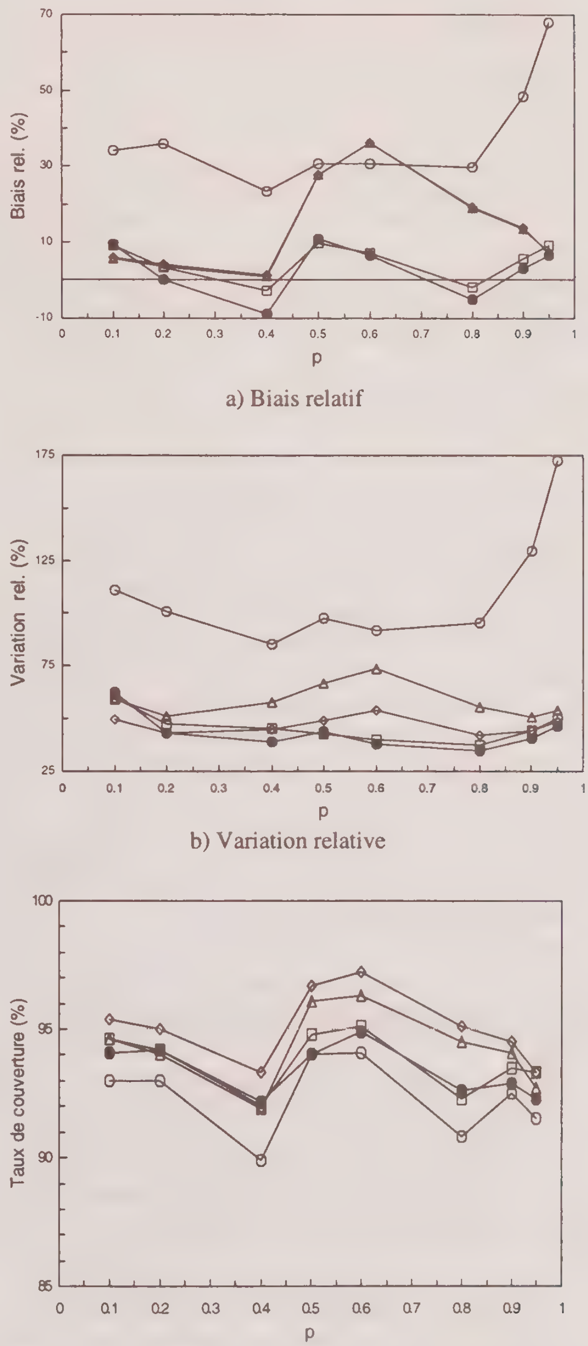
En ce qui concerne la stabilité des estimateurs de la variance, notre étude indique que les diverses méthodes donnent des résultats comparables, celle des équations d'estimation étant toutefois légèrement supérieure aux autres. Elle montre aussi qu'il existe une corrélation directe manifeste entre la mesure relative de la variation et la valeur de p.

Si nous comparons les méthodes d'après les propriétés de couverture des estimateurs de la variance de l'ordonnée de la courbe de Lorenz et de la part du quantile, nous constatons que, pour l'intervalle de confiance nominal de 95%, la méthode du jackknife produit des taux de couverture empiriques variant de 94.5% à 96.5% pour l'ordonnée de la courbe de Lorenz et de 94.5% à 99% pour la part du quantile. Les autres méthodes donnent des résultats comparables, les taux de couverture variant de 88% à 94%. Aussi bien dans le cas de l'ordonnée de la courbe de Lorenz que dans celui de la part du quantile, la couverture est meilleure quand la valeur de p est faible (voir figure 1c). Contrairement aux résultats obtenus pour l'indice de Gini, le taux d'erreur calculé pour l'intervalle latéral inférieur est environ deux fois plus grand que celui calculé pour l'intervalle latéral supérieur dans le cas de toutes les méthodes, pour l'ordonnée de la courbe de Lorenz ainsi que pour la part du quantile. Les résultats sont similaires pour les intervalles de confiance de 90% et de 99%.

Nos résultats empiriques donnent à penser que la méthode jackknife ne convient pas pour estimer la variance de l'ordonnée de la courbe de Lorenz ou de la part du quantile, particulièrement pour les grandes et les petites valeurs de p. Les méthodes du groupement de demi-échantillons compensés et du groupement répété de demi-échantillons compensés donnent de meilleurs résultats. Toutefois, le choix le plus judicieux est celui de la méthode des équations d'estimation ou de la méthode bootstrap.

Quantiles

Les résultats obtenus pour les quantiles sont présentés dans Kovačević, Yung et Pander (1995) et résumés graphiquement ici. Le biais relatif de l'estimation de la variance des quantiles selon la méthode jackknife varie de 23% à 67% pour l'estimateur JK1 et de 17% à 52% pour l'estimateur JK2. Les surestimations les plus importantes sont obtenues pour les variances de $\hat{\xi}_{0.90}$ et $\hat{\xi}_{0.95}$. Le tableau est nettement différent pour les méthodes GRDÉC et GDÉC. Tandis que la variance de la médiane calculée selon ces méthodes est surestimée de 27%, l'estimation de la variance des quantiles latéraux est très précise, le biais relatif variant de 3% à 7%. Les autres méthodes étudiées donnent également de beaucoup meilleurs résultats pour les quantiles latéraux et des résultats raisonnablement meilleurs pour la médiane et les quantiles qui l'entourent. Plus précisément, la méthode bootstrap et celle des équations d'estimation produisent les estimations entachées des biais relatifs les plus petits, sans que se dégage toutefois une tendance nette quant à la direction du biais. Pour les estimateurs bootstrap, le biais relatif se situe dans l'intervalle (−5%, +9%) et pour les équations d'estimation, dans l'intervalle (−8%, +9%) (voir le figure 3a).



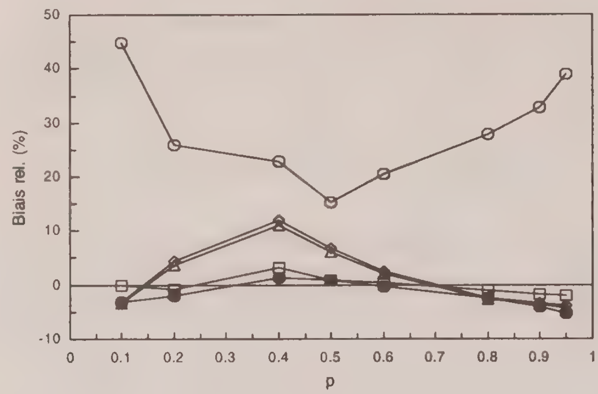
a) Biais relatif

b) Variation relative

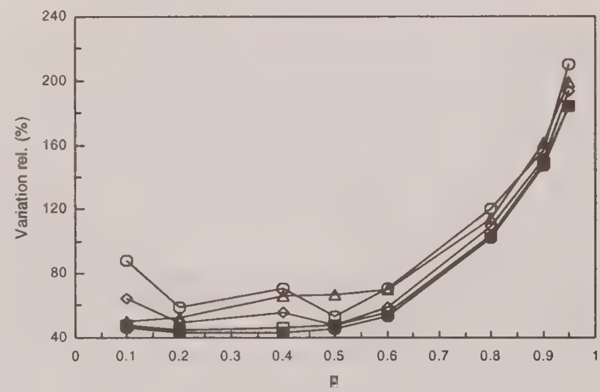
c) Taux de couverture (pour l'intervalle nominal de 95%)

○-JK1 ◇-GBHS1 △-RGBHS1 □-BS1 ●-ÉE

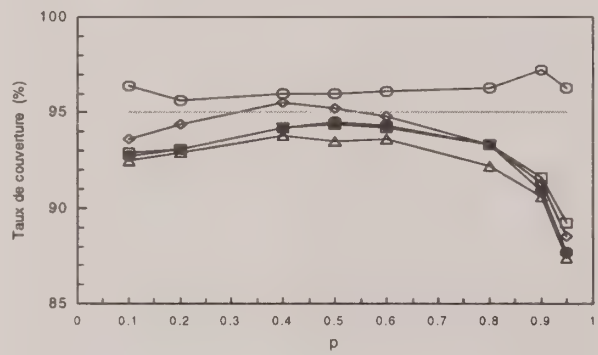
Figure 3. Propriétés des estimateurs de la variance des quantiles



a) Biais relatif



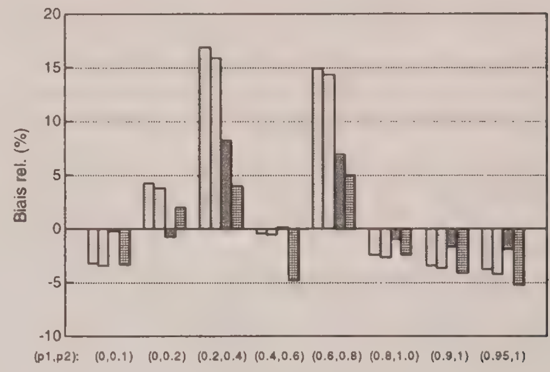
b) Variation relative



c) Taux de couverture (pour le taux nominal de 95%)

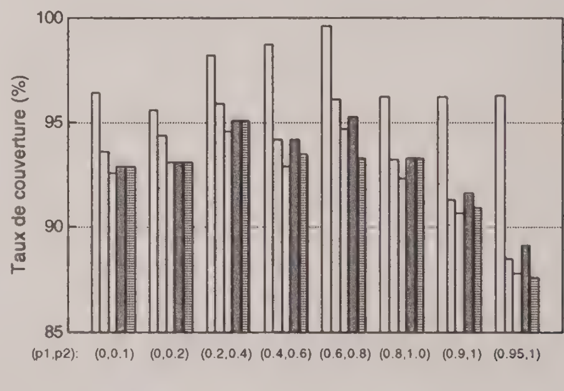
○-JK1 ◇-GBHS1 △-RGBHS1 □-BS1 ●-EE

Figure 1. Propriétés des estimateurs de la variance de l'ordonnée de la courbe de Lorenz



a) Biais relatif
(les estimateurs jackknife ne sont pas représentés)

b) Variation relative
(les estimateurs jackknife ne sont pas représentés)



c) Taux de couverture (pour l'intervalle nominal de 95%)

JK1 GBHS1 RGBHS1 BS1 EE

Figure 2: Propriétés des estimateurs de la variance de la part du quantile

Tableau 4
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de l'indice de polarisation

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		95.4	56.5	13.9	11.2	14.7	12.1	6	2.9	4.2
Variation relative (%)		138.7	78.5	77.5	75.9	60	58.6	48.4	47	50
Probabilité de couverture (95%)		98.6	98	94.2	93.8	95.4	95.2	95	94.7	94.4
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	0.7	0.8	2.2	2.4	1.4	1.4	1.8	2	2
	Inf.	0.8	1.1	3.6	3.9	3.2	3.4	3.2	3.4	3.6

parallèlement, les moins stables, de la variance. Les probabilités empiriques de couverture sont aussi très semblables pour tous les estimateurs. Les valeurs obtenues pour les taux d'erreur latéraux donnent à penser que l'utilisation d'intervalles de confiance asymétriques serait plus appropriée.

Proportion de faible revenu (PFR)

Toutes les méthodes étudiées ont tendance à produire une valeur surestimée de la variance de la PFR. Cependant, la grandeur de la surestimation varie fortement selon la méthode, l'écart allant de 1.1% pour la méthode des équations d'estimation (ÉE) à 76.9% pour la méthode jackknife (JK1). Des méthodes de rééchantillonnage, c'est la méthode bootstrap qui donne les meilleurs résultats, le biais relatif étant de 8.9% pour l'estimateur BS1 et de 3.8% pour l'estimateur BS2.

L'estimation de la variance de la proportion de faible revenu selon la méthode jackknife est très instable. Les estimateurs GRDÉC sont également moins stables. Les estimateurs bootstrap et ÉE donnent des résultats comparables, la variation relative allant de 31% à 45%.

Dans le cas de l'intervalle de confiance de 95%, les taux de couverture calculés d'après les estimations de la variance selon la méthode jackknife sont supérieurs aux taux nominal, soit 97.4% et 96.9%, à cause de la surestimation de la variance. Les autres méthodes produisent des taux de couverture un peu plus faibles que le taux nominal. Le calcul des taux d'erreur précisant les intervalles latéraux montre que toutes les méthodes aboutissent, pour l'intervalle inférieur, à un taux supérieur au taux nominal, ce qui laisse entendre que la proportion de faible revenu est caractérisée par une distribution asymétrique, fortement allongée vers la droite. Les taux de couverture et les taux d'erreur latéraux que nous obtenons pour les intervalles de confiance de 90% et de 99% suivent exactement la même tendance.

Dans l'ensemble, les méthodes bootstrap et ÉE sont nettement supérieures aux autres méthodes envisagées pour estimer la variance de la proportion de faible revenu.

Indice de polarisation

Les statistiques calculées pour évaluer les estimateurs de la variance de l'indice de polarisation indiquent que l'efficacité des diverses méthodes appliquées concordent fortement avec celle observée pour l'estimation de la variance de la proportion de faible revenu. De nouveau, ce sont les méthodes bootstrap et ÉE qui donnent les meilleurs résultats.

Ordonnées de la courbe de Lorenz et parts des quantiles

Les résultats complets obtenus pour les ordonnées de la courbe de Lorenz et pour les parts des quantiles figurent dans Kovačević, Yung et Pandher (1995). Nous présentons ici un résumé graphique des résultats (figures 1a à 1c). La méthode jackknife (les deux estimateurs) mène à une surestimation significative de la variance de toutes les ordonnées de la courbe de Lorenz (OCL) et de toutes les parts de quantile (PQ) envisagées. Pour l'OCL, le biais relatif de l'estimateur JK1 varie de 15% à 45% et celui de l'estimateur JK2, de 9% à 27%. Le biais relatif est plus faible au milieu de l'intervalle $(0 \le p \le 1)$ et presque trois fois plus grand dans les queues (pour les grandes et les petites valeurs de p). Le biais relatif de l'estimateur JK1 est pratiquement 50% plus grand que celui de l'estimateur JK2 pour l'OCL. L'écart est attribuable à la différence significative entre l'estimation de l'OCL à partir de l'échantillon complet, d'une part, et la moyenne calculée pour les itérations jackknife, d'autre part.

Nous obtenons des résultats similaires en ce qui concerne les estimateurs jackknife de la variance de la part du quantile, la variance étant surestimée de 26% à 237%, selon la part de population. La surestimation la plus importante se manifeste au milieu de l'intervalle. De nouveau, le biais relatif de l'estimateur JK1 surpasse d'environ 75% celui de l'estimateur JK2.

Le biais relatif observé pour les autres méthodes est très petit. Cependant, aucune tendance nette ne se dégage quant à la direction du biais – il est parfois positif, mais souvent négatif. La méthode bootstrap et la méthode des équations d'estimation donnent de meilleurs résultats que les autres, particulièrement autour de l'ordonnée de la courbe de Lorenz correspondant à $p = 0.5$ (voir figure 2a). Pour rendre l'illustration graphique plus claire, nous ne présentons pas les estimateurs jackknife dans les graphiques 2a et 2b.

Les résultats de l'estimation de la variance de la part du quantile sont similaires. La méthode bootstrap et celle des équations d'estimation sont celles qui fournissent les estimations les plus précises de la variance de l'ordonnée de la courbe de Lorenz et de la part du quantile. Pour l'ordonnée

Pour évaluer l'efficacité des intervalles de confiance fondés sur la loi normale, nous avons calculé les taux empiriques de couverture pour des niveaux de confiance nominaux de $100(1-\alpha)\% = 90, 95$ et 99%,

$$\text{cov. prob. } (v_M) = \frac{\sum_a I\{|\hat{\theta}_a - \theta| \big/ \sqrt{v_M(a)} \le z_{\alpha/2}\}}{A},$$

où $z_{\alpha/2}$ représente le $\alpha/2$-ième percentile centré réduit supérieur. Nous avons également calculé les taux d'erreur pour les intervalles latéraux supérieur et inférieur de la façon suivante,

$$\text{err\_}L(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \big/ \sqrt{v_M(a)} < -z_{\alpha/2}\}}{A}$$

$$\text{err\_}U(v_M) = \frac{\sum_a I\{(\hat{\theta}_a - \theta) \big/ \sqrt{v_M(a)} > z_{\alpha/2}\}}{A}.$$

Nous allons maintenant résumer le grand ensemble de résultats de la simulation séparément pour chaque mesure de l'inégalité du revenu.

4.2 Sommaire des résultats

Indice de Gini

Dans le cas de l'indice de Gini, la précision de toutes les méthodes d'estimation de la variance est comparable, le biais relatif, très faible, variant de −2.2% à −0.6%. De tous les estimateurs, c'est l'estimateur GRDÉC qui est entaché du biais relatif le plus faible.

Les divers estimateurs ont tous à peu près la même stabilité, de l'ordre de 87% à 99%. Les méthodes fondées sur le groupement de demi-échantillons compensés (GDÉC et GRDÉC) donnent des résultats un peu moins bons que les autres.

Pour l'intervalle de confiance de 95%, la probabilité de couverture varie de 92.6 (pour le GDÉC) à 93.9 (pour le GRDÉC). Le taux nominal de 2.5 % est une sous-estimation du taux d'erreur pour l'intervalle latéral inférieur dans le cas de toutes les méthodes étudiées. Selon nos résultats, le taux d'erreur pour l'intervalle inférieur varie de 4.6% à 5.4%, donc est plus de 100% plus élevé que le taux nominal de 2.5%. Inversement, le taux nominal est une surestimation du taux d'erreur pour l'intervalle supérieur dans le cas de toutes les méthodes (voir le tableau 2). Pour les intervalles de confiance de 90% et de 99%, les probabilités de couverture que nous avons calculées varient de 87.2 (pour le GDÉC) à 88.5 (pour le GRDÉC) et de 97.7 (pour le GDÉC) à 98.5 (pour le GRDÉC), respectivement. Pareillement, la comparaison des taux d'erreur obtenus pour les intervalles bilatéraux aux taux nominaux de 5% et de 1% indique la même tendance que pour le taux de 2.5%.

Dans l'ensemble, il est difficile de dire laquelle des méthodes d'estimation de la variance de l'indice de Gini étudiées ici est la meilleure, puisqu'elles produisent toutes des résultats comparables. L'application des méthodes à demi-échantillons compensés représente un léger compromis entre l'exactitude et la stabilité, puisque ces méthodes produisent les estimations les plus exactes, mais

Tableau 2
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de l'indice de Gini

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		−1.3	−1.3	−0.9	−1.1	−0.6	−0.7	−1.2	−2.2	−1.5
Variation relative (%)		87.1	87.1	99.4	99.2	95.2	95.1	88.5	87.6	87.0
Probabilité de couverture (95%)		93.8	93.8	92.6	92.6	93.9	93.9	93.5	93.4	93.7
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	4.8	4.8	5.4	5.4	4.6	4.6	5.0	5.1	4.9
	Inf.	1.4	1.4	2.0	2.0	1.5	1.5	1.5	1.5	1.4

Tableau 3
Valeurs des statistiques calculées pour évaluer les estimateurs de la variance de la proportion de faible revenu

		Jackknife		GDÉC		GRDÉC		Bootstrap		Équations d'estimation
		v_{J1}	v_{J2}	v_{GB1}	v_{GB2}	v_{RG1}	v_{RG2}	v_{B1}	v_{B2}	v_{EE}
Biais relatif (%)		76.9	58.4	25.8	21.0	26.8	21.9	8.9	3.8	1.1
Variation relative (%)		113.1	81.0	62.5	61.0	40.8	39.5	35.1	33.5	31.0
Probabilité de couverture (95%)		97.4	96.9	94.6	94.1	96.2	95.7	93.9	93.3	93.2
Taux d'erreur des intervalles latéraux (2.5%)	Sup.	2.1	2.6	3.3	3.5	2.4	2.6	4.6	5.0	5.0
	Inf.	0.5	0.6	2.0	2.4	1.4	1.7	1.5	1.7	1.7

Tableau 1
Définition de la variable aléatoire u^*_{hci} pour la méthode des équation d'estimation

Mesure	u^*_{hci}
Indice de Gini	$2[\hat{A}(y_{hci})y_{hci} + \hat{B}(y_{hci}) - \hat{\mu}(\hat{G}+1)/2]/\hat{\mu}$ où $A(y) = \hat{F}(y) - \frac{\hat{G}+1}{2}$ et $B(y) = \sum_s w_{hcj} y_{hcj} I\{y_{hcj} \geq y\}$.
Courbe de Lorenz	$[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)]/\hat{\mu}$
Part du quantile	$\frac{1}{\hat{\mu}}[(y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\hat{Q}(p_1, p_2)]$
Quantile	$-[I\{y \leq \hat{\xi}_p\} - p]/\hat{f}(\hat{\xi}_p)$, $\hat{f}(.)$ is the finite population density estimator
Proportion de faible revenu	$-\frac{\hat{f}(\hat{\xi}_{0.5}/2)}{2\hat{f}(\hat{\xi}_{0.5})}[I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 1/2] + [I\{y_{hci} \leq \hat{\xi}_{0.5}/2\} - \hat{\Lambda}_{0.5}]$
Indice de polarisation	$\frac{2}{\hat{\xi}_{0.5}}[(\hat{\xi}_{0.5} - y_{hci})(I\{y_i \leq \hat{\xi}_{0.5}\} - 0.5) - (A(y_{hci})y_{hci} + B(y_{hci}) - (\hat{G}+1)\hat{\xi}_{0.5}/2 + \hat{G}y_{hci}/2)] + \frac{P\hat{I}}{\hat{\xi}_{0.5}\hat{f}(\hat{\xi}_{0.5})}(I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) - P\hat{I}$

4. ÉTUDE EN SIMULATION

4.1 Données et conception de l'étude en simulation

La population sous-jacente à l'étude est celle de l'échantillon tiré en Ontario à l'occasion de l'Enquête canadienne sur les finances des consommateurs (EFC) de 1988. L'EFC est un complément annuel de l'Enquête mensuelle sur la population active. La population étudiée contient 7,474 ménages répartis entre 525 UPÉ tirées de 40 strates. Originellement, l'échantillon de l'Ontario a été sélectionné à partir de 91 strates que nous avons regroupées pour former des strates suffisamment grandes. Pour chaque ménage, nous disposions d'une valeur non négative du revenu annuel total. La courbe de répartition du revenu obtenue pour cette micro-population était fortement étalée vers la droite, les coefficients d'asymétrie et d'aplatissement étant égaux à 4.5 et 89.5, respectivement. Nous avons calculé les valeurs réelles des paramètres étudiés (mesures de l'inégalité et de la polarisation du revenu) pour cette population. Suivant la méthode de Neyman, nous avons répartit 108 grappes (UPÉ) entre les 40 strates. Puis, selon un plan d'échantillonnage par grappes à un degré, nous avons sélectionné avec probabilité proportionnelle à la taille et avec remise, des échantillons de strate dont la taille variait de 2 à 6 grappes. Enfin, dans une grappe sélectionnée, nous avons dénombré tous les ménages (6 à 20).

Les mesures du revenu sur lesquelles porte l'étude sont l'indice de Gini, la proportion de faible revenu, l'indice de polarisation, un ensemble de parts du quantile, un ensemble d'ordonnées de la courbe de Lorenz et les quantiles correspondants. Nous avons utilisé comme approximation de l'erreur quadratique moyenne des estimations de ces mesures la valeur de l'erreur quadratique moyenne empirique (EMSE pour *empirical mean squared error*), calculée sur 10,000 échantillons indépendants tirés conformément au plan d'échantillonnage décrit plus haut. Nous avons utilisé ces EMSE en remplacement des erreurs quadratiques moyennes «réelles» pour la comparaison avec les variances estimées.

Pour chacun des 10,000 échantillons, en plus de l'estimation des paramètres, nous avons estimé la variance d'échantillonnage en appliquant diverses méthodes, à savoir la méthode jackknife avec suppression d'une UPÉ (JK), la méthode du groupement de demi-échantillons compensés (GDÉC) et celle du groupement répété de demi-échantillons compensés (GRDÉC), la méthode bootstrap (BS) et la méthode de linéarisation basée sur des équations d'estimation (ÉE). Pour toutes les méthodes de rééchantillonnage, nous nous sommes servis de deux estimateurs distincts, l'un correspondant à l'estimation d'après l'échantillon complet et l'autre à la moyenne de toutes les répétitions. Le calcul des estimateurs de la variance est fondé sur 108 itérations dans le cas de la méthode jackknife et sur 100 itérations dans le cas de la méthode bootstrap. Les méthodes GDÉC et GRDÉC se fondent sur 44 échantillons répétés compensés obtenus à partir d'une matrice de Hadamard de 44 sur 44, avec en plus 3 répétitions pour la méthode GRDÉC, ce qui donne en tout, dans ce cas, 132 demi-échantillons répétés. Il convient de souligner que le nombre d'itérations jackknife dépend du nombre de grappes dans l'échantillon, donc n'est pas arbitraire. Pareillement, dans le cas de la méthode GDÉC, le nombre de strates détermine le nombre d'échantillons répétés. Pour que le nombre d'échantillons répétés soit comparable pour toutes les méthodes, nous avons décidé d'effectuer 100 ($\approx$ 108) itérations bootstrap et 3 répétitions du GDÉC pour obtenir 132 échantillons répétés dans le cas du GRDÉC.

Afin d'évaluer l'exactitude et la précision des méthodes étudiées, nous avons calculé leur biais relatif (*rel. bias*) et leur variance relative (*rel. var.*), c'est-à-dire leur instabilité, sur les $A = 10{,}000$ simulations, soit

$$\text{rel. bias}(v_M) = \frac{\sum_a v_M(a)/A - \text{EMSE}}{\text{EMSE}}$$

$$\text{rel. var.}(v_M) = \frac{\sqrt{\sum_a [v_M(a) - \text{EMSE}]^2/A}}{\text{EMSE}}$$

variance selon la méthode GDÉC concorde avec un estimateur convergent de la variance dans le cas des statistiques linéaires. Donc, nous établissons pour les facteurs de perturbation $1 - a_h$: particuliers aux strates les contraintes suivantes:

Pour toutes les valeurs de h: (i) $0 < 1 - a_h \leq 1$; (ii) $(1 - a_h)^2 (m_{h_1}/m_{h_2})^2 \approx 1$; (iii) $(1 - a_h)^2 m_{h_1}/m_{h_2} \approx 1$. Pour les n_h pairs, nous posons simplement $1 - a_h = 1$. Cependant, maintenir la contrainte $1 - a_h = 1$ pour les valeurs impaires de n_h exclurait toute contribution des grappes du premier demi-échantillon quand $\delta_h^{(r)} = -1$, [voir l'équation (3.4)]. Aux fins de l'étude en simulation, nous choisissons

$$1 - a_h = \sqrt{\frac{n_h}{2\, m_{h2}}} \tag{3.5}$$

qui se réduit à 1 quand n_h est pair. Quand la taille de l'échantillon de strate est impaire, l'expression est égale à $\sqrt{1 - 1/(n_h + 1)}$. Dans notre simulation, très peu de strates ont un n_h impair et nous obtenons $v_{GB1}(\hat{\mu}_Y) = v_{GB2}(\hat{\mu}_Y) \approx v_L(\hat{\mu}_Y)$, où $\hat{\mu}_Y$ est la moyenne de l'échantillon et où $v_L(\hat{\mu}_Y)$ est l'estimateur de la variance de linéarisation utilisé couramment. Cependant, nous sommes d'avis qu'il faut poursuivre les travaux visant à modifier la méthode GDÉC de façon à pouvoir traiter un grand nombre de strates contenant un nombre impair d'UPÉ.

À l'instar de la méthode jackknife, l'estimation de la mesure de l'inégalité du revenu calculée à partir du r-ième demi-échantillon est donné par $\hat{\theta}^{(r)} = \sum_s J(\hat{F}^{(r)}, y_{hci}, \hat{\beta}^{(r)}) \tilde{w}_{hci}^{(r)}$ où $\hat{\beta}^{(r)}$ est une estimation du paramètre dérangeant fondée sur le r-ième demi-échantillon et où $\tilde{w}_{hci}^{(r)} = \tilde{w}_{hci} A_{hc}^{(r)}$. L'estimateur résultant de la variance de $\hat{\theta}$ selon la méthode GDÉC prend la forme

$$v_{GB1}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.6}$$

Si nous répétons T fois le groupement aléatoire des unités dans chaque strate, que nous calculons chaque fois $v_{GB1}(\hat{\theta})$ et que nous calculons la moyenne pour les T répétitions, nous obtenons l'estimateur de la variance selon la méthode du groupement répété des demi-échantillons compensés (GRDÉC), soit

$$v_{RG1}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} v_{GB1_t}(\hat{\theta}).$$

En remplaçant $\hat{\theta}$ par $\hat{\theta} = \sum_r \hat{\theta}^{(r)}/R$, nous obtenons une variante de l'estimateur GDÉC (et GRDÉC) que nous dénotons $v_{GB2}(\hat{\theta})$ (et $v_{RG2}(\hat{\theta})$).

Inutile de dire que, pour étalonner les poids, il faut les modifier comme il convient pour chaque répétition GDÉC selon la même méthode avec demi-échantillon compensé.

3.3 Méthode bootstrap

Nous avons également examiné les résultats obtenus en appliquant la méthode bootstrap pour estimer la variance de diverses statistiques du revenu. Nous avons adopté pour l'échantillon en grappes stratifié le plan de rééchantillonnage bootstrap préconisé par Rao, Wu et Yue (1992). Brièvement, il s'agit de sélectionner indépendamment dans chaque strate un échantillon aléatoire simple de $n_h - 1$ grappes avec remise (pour les n_h grappes). On obtient le poids bootstrap, w_{hci}^*, en modifiant le poids w_{hci} original de la façon suivante:

$$w_{hci}^* = A_{hc} w_{hci}$$

où

$$A_{hc} = \frac{n_h}{n_h - 1} m_{hc}^*$$

et où m_{hc}^* est le nombre de fois que la hc-ième grappe est sélectionnée. Il convient de souligner que $\sum_c m_{hc}^* = n_h - 1$. La méthode est répétée indépendamment B fois; pour chaque échantillon bootstrap, nous calculons $\hat{\theta}^* = \sum_s J(\hat{F}^*, y_{hci}, \hat{\beta}^*) \tilde{w}_{hci}^*$, où $\hat{\beta}^*$ est une estimation du paramètre dérangeant fondée sur l'échantillon bootstrap et où $\tilde{w}_{hci}^* = w_{hci}^* / \sum_s w_{hci}^*$. L'estimation bootstrap de la variance de $\hat{\theta}$ est donnée par

$$v_{B1}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{(b)}^* - \hat{\theta})^2.$$

On obtient une autre estimation de la variance en remplaçant $\hat{\theta}$ par la moyenne des répétitions bootstrap.

3.4 Linéarisation par la méthode des équations d'estimation

Contrairement aux méthodes de rééchantillonnage, la méthode des équations d'estimation (ÉE) de Binder (Binder, 1991; Binder et Patak 1994; Binder et Kovačević 1995), ne demande pas énormément de calcul. Cette méthode, fondée sur la linéarisation, produit des formules de variance asymptotique faciles à programmer malgré leur apparence complexe.

En appliquant les méthodes des équations d'estimation décrites par Binder et Patak (1994), Binder et Kovačević (1995) et Kovačević et Binder (1997), nous obtenons pour les estimateurs de la variance approximative des mesures étudiées du revenu des expressions telles que

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1} \sum_c \left(u_{hc}^* - \bar{u}_h^* \right)^2 \tag{3.7}$$

où, $u_{hc}^* = \sum_i \tilde{w}_{hci} u_{hci}^*$, $\bar{u}_h^* = \sum_c u_{hc}^* / n_h$ et $\tilde{w}_{hci}$ est un poids normalisé. Pour obtenir plus de précisions sur la méthode des équations d'estimation, en particulier sur la relation entre les variables aléatoires u_{hci}^* et la fonction J, le lecteur devrait consulter Binder et Kovačević (1995). Pour les mesures du revenu étudiées ici, les variables aléatoires u_{hci}^* figurent dans le tableau 1.

Dans les cas de la proportion de faible revenu et de l'indice de polarisation, les expressions représentant les variables aléatoires u_{hci}^* dépendent de l'estimation de la fonction de densité correspondant à la médiane, $\hat{f}(\hat{\xi}_{0.5})$, et à la moitié de la médiane, $\hat{f}(\hat{\xi}_{0.5}/2)$. Binder et Kovačević (1995) décrivent une méthode appropriée pour estimer ces quantités.

où $\hat{\boldsymbol{\beta}}$ représente le vecteur estimé des paramètres dérangeants et où $\tilde{w}_{hci}$ sont les poids normalisés. Selon cette formule générale, l'estimation d'une mesure de l'inégalité du revenu calculée d'après l'échantillon après suppression de gj UPÉ est donnée par

$$\hat{\theta}_{(gj)} = \sum_s J(\hat{F}_{(gj)}, y_{hci}, \hat{\boldsymbol{\beta}}_{(gj)})\, \tilde{w}_{hci(gj)}$$

où $\hat{F}_{(gj)}$ et $\hat{\boldsymbol{\beta}}_{(gj)}$ sont les valeurs de la fonction de distribution et du paramètre dérangeant estimées d'après l'échantillon dont on a supprimé la gj-ième UPÉ et où

$$\tilde{w}_{hci(gj)} = \begin{cases} w_{hci}/\hat{N}_{(gj)}, & \text{si} \quad h \neq g, \\ \dfrac{n_g}{n_g - 1} w_{gci}/\hat{N}_{(gj)}, & \text{si} \quad h = g, c \neq j, \\ 0, & \text{si} \quad h = g,\ c = j. \end{cases}$$

L'estimateur jackknife «avec suppression d'une UPÉ» résultant de la variance $\hat{\theta}$ est

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2 . \tag{3.2}$$

Si on remplace $\hat{\theta}$ par $\hat{\theta}_{..} = \sum_g \sum_j \hat{\theta}_{(gj)}/n$, on obtient une variante de l'estimation jackknife de la variance. Nous la représentons par $v_{J2}(\hat{\theta})$. Manifestement, $v_{J2}(\hat{\theta}) \leq v_{J1}(\hat{\theta})$. Krewski et Rao (1981) ont démontré la convergence de (3.2) pour les statistiques lissées.

Dans le cas de l'estimation de la variance des quantiles et de la fonction de quantiles, nous commençons par calculer les quantiles d'après l'échantillon dont on a supprimé la gj-ième UPÉ,

$$\hat{\xi}_{(gj)}(p) = \inf \{ y_{hci} \mid \hat{F}_{(gj)}(y_{hci}) \geq p, hci \in s \setminus (gj) \},$$

puis nous calculons $\hat{\theta}_{(gj)} = g(\hat{\xi}_{(gj)})$, et enfin nous utilisons l'équation (3.2) pour obtenir l'estimateur jackknife de la variance.

3.2 Méthode du groupement de demi-échantillons compensés (GDÉC) et méthode du groupement répété de demi-échantillons compensés (GRDÉC)

Originellement, la méthode des demi-échantillons compensés a été proposée pour les plans de sondage produisant deux grappes par strate. Or, le cas qui nous intéresse est celui où les strates comptent plus de deux grappes. Ordinairement, dans cette situation, on répartit les grappes (unités de premier degré) de chaque strate en deux groupes. Nous explorons l'idée formulée par Wu (1991) et nous simplifions son application en vue d'estimer la variance de la FDC. En premier lieu, dans chaque strate $h, (h = 1, \ldots, L)$, nous groupons les UPÉ au hasard en deux moitiés, h_1 et h_2, contenant $m_{h_1} = [n_h/2]$ et $m_{h_2} = n_h - m_{h_1}$ UPÉ, respectivement. Si nous posons que l'indicateur de groupe est

$$\delta_h^{(r)} = \begin{cases} 1, & h_1 \in r \\ -1, & h_2 \in r \end{cases}$$

où $r = 1, \ldots, R$ représente un demi-échantillon (échantillon répété), les demi-échantillons sont compensés sur les groupes si $\sum_{r=1}^{R} \delta_h^{(r)} = 0$ et $\sum_{r=1}^{R} \delta_h^{(r)} \delta_{h'}^{(r)} = 0, (h \neq h')$. L'utilisation d'une matrice de Hadamard d'ordre $R\,(L + 1 \leq R \leq L + 4)$ permet d'obtenir un ensemble minimal de demi-échantillons compensés.

L'estimateur de la fonction de distribution fondé sur le r-ième demi-échantillon est

$$\hat{F}^{(r)}(y) = \frac{\hat{G}^{(r)}(y)}{\hat{N}^{(r)}}$$

où

$$\hat{G}^{(r)}(y) = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}\, I\{ y_{hci} \leq y \}, \hat{N}^{(r)} = \sum_h \sum_c A_{hc}^{(r)} \sum_i w_{hci}$$

et où $A_{hc}^{(r)}$ représente le modificateur de poids qui demeure constant pour toutes les grappes d'un même demi-échantillon. Nous supposons que les poids de toutes les unités (ménages) d'une grappe sont rééchelonnés de façon égale par le modificateur $A_{hc}^{(r)}$.

Dans le cas de la méthode GDÉC type, quand n_h est pair, nous utilisons

$$A_{hc}^{(r)} = \begin{cases} 1 + \delta_h^{(r)}, & c \in h_1, \\ 1 - \delta_h^{(r)}, & c \in h_2 \end{cases} \tag{3.3}$$

ce qui signifie que les poids sont modifiés par un facteur 2 ou un facteur 0 selon qu'une unité figure ou non dans l'échantillon répété. Dans les cas où n_h est impair, diverses modifications ont été envisagées [consulter Shao (1993) et Sitter (1993)].

La méthode que nous appliquons se fonde sur le plan type de rééchantillonnage avec répétition compensée et sur une variante de la méthode de rééchelonnage proposée par Shao (1993), soit

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - b_h)\delta_h^{(r)}, & c \in h_2. \end{cases}$$

Le fait de maintenir la taille de l'échantillon de strate dans tout demi-échantillon répété signifie que

$$\sum_{c \in h1} [1 + (1 - a_h)\delta_h^{(r)}] + \sum_{c \in h2} [1 - (1 - b_h)\delta_h^{(r)}] = n_h,$$

ce qui donne

$$A_{hc}^{(r)} = \begin{cases} 1 + (1 - a_h)\, \delta_h^{(r)}, & c \in h_1; \\ 1 - (1 - a_h) \dfrac{m_{h_1}}{m_{h_2}} \delta_h^{(r)}, & c \in h_2. \end{cases} \tag{3.4}$$

Pour être certain que les poids modifiés ne soient pas négatifs, a_h devrait satisfaire $0 \leq a_h < 1$. Quand n_h est pair, nous aimerions que (3.4) se réduise à (3.3). Conformément à l'idée de Shao (1993), nous voulons que l'estimateur de la

(1986)). Rao et Wu (1988) proposent une méthode bootstrap modifiée pour l'estimation de la variance dans le cas de plans d'échantillonnage stratifiés. Kovar (1987) et Kovar et coll. (1988) indiquent que la méthode donne de bons résultats pour la médiane quand la taille de l'échantillon auquel on l'applique est $n_h^* = n_h - 1$.

Dans le cas de la méthode d'estimation de la variance par groupement de demi-échantillons compensés (GDÉC), les grappes échantillonnées dans chaque strate sont réparties au hasard en deux groupes (demi-échantillons) auxquels on applique la méthode de répétition compensée. Rao et Shao (1996) montrent que cette méthode est asymptotiquement incorrecte en ce sens que la distribution du facteur pivot t associé ne converge pas vers une distribution normale réduite et que les intervalles de confiance connexes sont asymptotiquement incorrects. Pour surmonter cette difficulté, ils proposent de répéter indépendamment le groupement T fois puis de calculer la moyenne des T estimations résultantes de la variance. Il montre qu'un tel estimateur est asymptotiquement correct pour un plan d'échantillonnage aléatoire stratifié quand $\min n_h \to \infty$ et $T \to \infty$. Grâce à une petite simulation, ils montrent que la méthode donne de bons résultats pour des valeurs de T aussi petites que 15 dans le cas d'estimateurs lissés. La méthode du groupement répété de demi-échantillons compensées (GRDÉC) produit un meilleur estimateur de la variance de la médiane de la population que les méthodes jackknife et GDÉC en ce sens que le biais relatif et le coefficient de variation sont plus faibles. Récemment, McCarthy (1993) a examiné et comparé diverses méthodes d'estimation de la variance de la médiane fondées sur l'échantillonnage aléatoire simple sans remise d'une population finie. Son étude englobe la plupart des méthodes de rééchantillonnage.

Bien que les méthodes de linéarisation, qui sont précieuses en statistiques non linéaire, soient difficiles à appliquer aux quantiles, puisqu'il faut estimer la densité, Binder (1991), Binder et Kovačević (1995) et Kovačević et Binder (1997) ont obtenu des estimateurs convergents de la variance de certaines mesures non lissées de l'inégalité et de la polarisation du revenu en appliquant une méthode de linéarisation à des équations d'estimation. Le calcul des estimateurs obtenus selon cette méthode est plus simple que celui des estimateurs fondés sur le rééchantillonnage, mais nécessite un calcul théorique.

Plusieurs auteurs ont étudié l'estimation de la variance de l'indice de Gini en supposant que les conditions d'échantillonnage aléatoires simples étaient respectées (Glasser 1962; Sendler 1979; Sandström, Wretman et Walden 1985; et Yitzhaki 1991). Dans le cas d'un plan d'échantillonnage complexe, Love et Wolfson (1976) proposent une méthode «grossière de répétition de demi-échantillons». Sandström, Wretman et Walden (1988) comparent les méthodes de calcul de la variance approximative à la méthode jackknife avec suppression d'une unité d'échantillonnage dans le cas de trois plans d'échantillonnage, dont deux complexes.

L'estimation de la variance des ordonnées de la courbe de Lorenz et des parts de quantile correspondantes a reçu moins d'attention. Le calcul de leur variance asymptotique est assez complexe. À cet égard, on mentionnera les percées effectuées par Beach et Davidson (1983) et par Beach et Kaliski (1986). Leurs travaux ont pour cadre de référence le modèle de la superpopulation en vertu duquel on considère les poids de sondage constants quand on construit les estimations. Le fait que cette méthode soit fondée sur un modèle pourrait restreindre son application à des données tirées d'enquêtes par sondage pour lesquelles on juge le plan d'échantillonnage significatif.

Dans les sous-sections qui suivent, nous passons en revue les méthodes d'estimation de la variance utilisées dans le cadre de la présente étude.

3.1 Méthode Jackknife avec suppression d'une UPÉ

La méthode repose sur l'exclusion séquentielle (suppression) d'une UPÉ à la fois du calcul de l'estimation. Après la suppression, on modifie les poids des unités retenues dans l'échantillon de sorte que les poids supprimés soient compensés et que la FDC estimée à partir de l'échantillon résiduel ait les mêmes propriétés que la FDC originale. Représentons par $\hat{F}_{(gj)}(y)$ l'estimation de la FDC fondée sur un échantillon dont on a supprimé la gj-ième UPÉ, c'est-à-dire

$$\hat{F}_{(gj)}(y) = \hat{G}_{(gj)}(y) / \hat{N}_{(gj)}$$

où

$$\hat{G}_{(gj)}(y) = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} I\{y_{hci} \leq y\} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci} I\{y_{gci} \leq y\}$$

et

$$\hat{N}_{(gj)} = \sum_{h \neq g} \sum_{c} \sum_{i} w_{hci} + \frac{n_g}{n_g - 1} \sum_{c \neq j} \sum_{i} w_{gci}.$$

L'estimateur jackknife «avec suppression d'une UPÉ» de la variance de $\hat{F}(y)$ est

$$v_{J1}(\hat{F}(y)) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{F}_{(gj)}(y) - \hat{F}(y))^2.$$

On peut établir la convergence asymptotique de $v_{J1}(\hat{F}(y))$ en se servant des résultats de Krewski et de Rao (1981).

Par souci de commodité, nous notons que l'on peut écrire toutes les mesures considérées ici sous la forme générale

$$\theta_N = \sum_{U} J(F_N, Y_i, \boldsymbol{\beta}) \frac{1}{N},$$

où $J(\cdot)$ est une fonction à valeur réelle qui dépend éventuellement du paramètre dérangeant $\boldsymbol{\beta}$. Le paramètre de population finie θ_N est alors estimé par

$$\hat{\theta} = \sum\nolimits_s J(\hat{F}, y_{hci}, \hat{\boldsymbol{\beta}})\, \tilde{w}_{hci} \tag{3.1}$$

Pour $0 \leq p_1 < p_2 \leq 1$ nous estimons la part du quantile en remplaçant les paramètres par leur valeur estimée.

La mesure la plus utilisée de l'inégalité globale de la répartition du revenu, à savoir l'*indice de Gini*, correspond, par définition, à la surface comprise entre la courbe de Lorenz et l'axe de 45°, normalisée de façon à ce que sa valeur soit comprise entre 0 et 1, c'est-à-dire $G = 1 - 2\int_0^1 L(p)\,dp$. Dans le cas d'une population finie, l'estimation de l'indice de Gini prend la forme

$$\hat{G} = \sum_s \frac{[2\hat{F}_{hci} - 1]y_{hci}}{\hat{\mu}}\tilde{w}_{hci}.$$

Le lecteur qui souhaite obtenir plus de précisions sur l'indice de Gini devrait consulter Nygärd et Sandström (1985).

De façon analogue à la courbe de Lorenz et à l'indice de Gini, Foster et Wolfson (1992) définissent la *courbe de polarisation* par la relation

$$B(p) = \int_{0.5}^{p} \frac{F^{-1}(q) - \xi_{0.5}}{\xi_{0.5}}\,dq,$$

qui, dans le cas d'une population finie, prend la forme

$$B(p) = \begin{cases} 0.5 - p - \dfrac{1}{\xi_{0.5}} \displaystyle\sum_U I\{p < F_i < 0.5\}\, Y_i \frac{1}{N}, & 0 < p \leq 0.5, \\ 0.5 - p + \dfrac{1}{\xi_{0.5}} \displaystyle\sum_U I\{0.5 \leq F_i < p\}\, Y_i \frac{1}{N}, & 0.5 < p \leq 1. \end{cases}$$

La courbe de polarisation montre, pour tout percentile de population, la mesure dans laquelle le revenu s'écarte de la médiane. La surface située au-dessous de la courbe de polarisation donne une mesure sommaire de la polarisation. La version normalisée, de façon à ce que sa valeur soit comprise entre 0 et 1, appelée *indice de polarisation* (*IP*), prend la forme

$$IP_N = \sum_U \frac{[2 - 2I\{F_i \leq 0.5\} - 2F_i]Y_i}{\xi_N(0.5)} \frac{1}{N}$$

où $\xi_N(0.5), \mu_N, F_i$ ont été définis plus haut. On obtient l'estimation de l'indice de polarisation en remplaçant les paramètres par leurs valeurs estimées dans cette équation.

3. ESTIMATION DE LA VARIANCE

L'estimation de la variance de statistiques non lissées, comme les quantiles, et des fonctions axées sur les quantiles, comme la proportion de faible revenu ou l'indice de polarisation, n'est pas chose facile, surtout quand on ne peut soutenir l'hypothèse de l'échantillonnage aléatoire simple et qu'il faut tenir compte d'un plan de sondage complexe. Dans la première partie de la présente section, nous passerons en revue certains résultats de l'estimation de la variance des quantiles qui permettront de mieux comprendre ultérieurement l'estimation complexe de la variance des mesures de l'inégalité du revenu, ainsi que les résultats de l'estimation de la variance de certaines mesures, dont les ordonnées de la courbe de Lorenz. Dans la deuxième partie, nous décrivons les méthodes d'estimation de la variance faisant l'objet de la présente étude.

Woodruff (1952) a proposé une méthode permettant de calculer les intervalles de confiance de quantiles distincts. À partir de ces intervalles, Francisco et Fuller (1986) et Rao et Wu (1987) ont établi les estimateurs de la variance. Bien que l'estimateur dépende du niveau de confiance, Rao et Wu (1987) ont déterminé sa convergence asymptotique pour tout seuil de signification α. Au moyen de simulations de Monte Carlo, ils ont étudié les écarts-types des quantiles des échantillons en grappes estimés. Leurs résultats donnent à penser que le choix d'un intervalle de confiance de 95% pour déduire l'écart-type donne de bons résultats. Binder (1991) obtient une forme comparable de l'estimateur de la variance en appliquant la méthode de linéarisation.

L'augmentation de la puissance informatique a rendu très courante l'utilisation d'estimateurs jackknife de la variance dans le cas des fonctions lissées des totaux et des moyennes. L'application de la théorie asymptotique type à la médiane d'une distribution à densité continue bornée, f, montre que $nE(\hat{\xi}_{0.5} - \xi_{0.5})^2 \to 1/[4f^2(\xi_{0.5})]$ quand $n \to \infty$. Efron (1979) fait remarquer que l'application de la méthode jackknife à la médiane de l'échantillon produit une estimation de la variance asymptotiquement non convergente puisque

$$n\,\mathrm{var}_{JK}(\hat{\xi}_{0.5}) \to \frac{1}{4f^2(\xi_{0.5})}[\chi_2^2/2]^2$$

où la moyenne de $[\chi_2^2/2]^2$ est égale à 2 et sa variance est égale à 20, ce qui signifie que l'estimateur jackknife de la variance a tendance à surestimer par 100%, en moyenne, la variance asymptotique correcte. Kovar (1987) confirme empiriquement l'incohérence des estimateurs jackknife avec suppression d'une unité pour un plan d'échantillonnage stratifié. Au moyen d'une simulation portant sur une population stratifiée, il montre que les estimateurs jackknife avec suppression d'une unité (il en examine six) donnent de piètres résultats, surestimant la variance réelle de 30 à 70% dans le cas d'un plan comptant deux unités par strate et des résultats encore pires dans le cas d'un plan comptant cinq unités par strate. Néanmoins, Shao et Wu (1989) montrent que, dans certaines conditions, la méthode jackknife avec suppression d'une unité présente des propriétés asymptotiques désirables pour l'estimation de la variance de statistiques non lissées. Ce résultat a encouragé Rao, Wu et Yue (1992) à appliquer la méthode jackknife avec suppression d'une UPÉ à l'échantillonnage stratifié à plusieurs degrés. Grâce à une étude en simulation limitée, il montre que le biais ainsi que le biais relatif de l'estimateur de la variance de la médiane diminuent à mesure que la taille de la grappe augmente quand la corrélation à l'intérieur de la grappe est invariable.

La méthode bootstrap d'estimation de la variance, mentionnée pour la première fois par Efron (1979), donne des résultats convergents dans le cas d'observations indépendantes, distribuées de façon identique (consulter aussi Babu

2. ESTIMATION DES MESURES DE L'INÉGALITÉ DU REVENU

Les moyens les plus simples de mesurer l'inégalité entre deux distributions consistent à comparer les fonctions de distribution cumulatives (FDC) ou les quantiles de ces distributions. Nous commencerons par définir la FDC et les quantiles d'une population finie. Les autres mesures examinées dans le présent article, qui sont des fonctions de la FDC ou un nombre déterminé de quantiles, seront présentées à la section 2.1.

Soit une variable Y de la population finie $U = \{1, ..., N\}$. Nous définissons la FDC de cette variable comme étant

$$F_N(y) = \sum_{i \in U} I\{Y_i \le y\} \frac{1}{N},$$

où $I\{a\}$ est une fonction indicatrice dont la valeur est 1 si a est vrai et 0 autrement. Nous représentons l'estimateur sans biais dû au plan de sondage de $F_N(y)$ par

$$\tilde{F}(y) = \sum_{i \in s} I\{y_i \le y\} \frac{w_i}{N}$$

où les poids d'échantillonnage, w_i, sont calculés d'après le plan de sondage et sont égaux à l'inverse des probabilités d'inclusion de premier ordre. Toutefois, cet estimateur pourrait ne pas être une FDC, puisque $\tilde{F}(\infty) = \hat{N}/N$ n'est pas nécessairement égale à 1. Donc, nous préférons utiliser l'estimateur éventuellement entaché d'un biais dû au plan de sondage, soit:

$$\hat{F}(y) = \sum_{i \in s} I\{y_i \le y\} w_i \Big/ \sum_{i \in s} w_i = \sum_{i \in s} I\{y_i \le y\} \tilde{w}_i, \qquad (2.1)$$

où $\tilde{w}_i = w_i / \sum_s w_i$, $i \in s$. L'estimateur (2.1) n'est entaché d'aucun biais dû au plan de sondage quand $\sum_s w_i = N$, autrement dit si on effectue un échantillonnage aléatoire simple ou qu'on étalonne les poids, w_i, d'après des totaux connus de la population. En général, on applique à l'estimateur (2.1) des poids finals, ordinairement obtenus après stratification à posteriori, correction pour la non-réponse, certains étalonnages itératifs, etc. Dans le présent article, nous examinons uniquement le cas où les poids sont établis d'après le plan de sondage.

Passons maintenant aux quantiles. Nous représentons les quantiles d'une population finie par la fonction

$$\xi_N(p) = \inf_{i \in U} \{Y_i \mid F_i \ge p\} \quad \text{pour} \quad 0 < p \le 1,$$

où $F_i = F_N(Y_i)$. Nous estimons les quantiles de la population d'après les quantiles de l'échantillon, soit

$$\hat{\xi}_p = \inf_{i \in s} \{y_i \mid \hat{F}_i \ge p\} \quad \text{pour} \quad 0 < p \le 1,$$

où $\hat{F}_i = \hat{F}(y_i)$. Si un paramètre est une fonction des quantiles, disons $\theta_N = g(\xi_N)$ avec $\xi_N = \{\xi_N(p_1), ..., \xi_N(p_k)\}$, alors son estimateur est $\hat{\theta} = g(\hat{\xi})$ où $\hat{\xi} = (\hat{\xi}_{p_1}, ..., \hat{\xi}_{p_k})$.

2.1 Mesures de l'inégalité et de la polarisation du revenu en tant que paramètres d'une population finie

Dans la présente section, nous présentons certaines mesures de l'inégalité et de la polarisation du revenu utilisées fréquemment, à savoir le seuil de faible revenu, la proportion de faible revenu, la courbe de Lorenz et les statistiques connexes, les parts des quantiles, l'indice de Gini et, enfin, la courbe de polarisation et l'indice de polarisation. Nous nous limiterons ici à les présenter brièvement. Pour plus de précisions, le lecteur consultera Nygärd et Sandström (1981) et Wolfson (1994).

Nous définissons le *seuil de faible revenu*, ou *seuil de pauvreté*, comme une fonction de la médiane, $\lambda_\alpha = \alpha\, \xi_N(0.5)$, où $0 < \alpha \le 1$ est une constante donnée et où $\xi_N(0.5)$ est la médiane de la population finie. Son estimation est simplement $\hat{\lambda}_\alpha = \alpha\, \hat{\xi}_{0.5}$.

La *proportion de faible revenu (PFR)* est la proportion d'unités (particuliers, familles, ménages) de la population qui se situe sous le seuil de faible revenu λ_α et est représentée par $\Lambda_\alpha = F_N(\lambda_\alpha)$. L'estimateur de la proportion de faible revenu englobe à la fois l'estimateur de la fonction de distribution et celui du seuil de faible revenu, $\hat{\Lambda}_\alpha = \hat{F}(\hat{\lambda}_\alpha) = \sum_s I\{y_{hci} \le \alpha\, \hat{\xi}_{0.5}\} \tilde{w}_{hci}$.

L'*ordonnée de la courbe de Lorenz (OCL)* d'une population finie précise la part du revenu que reçoit le percentile $100p$ le plus pauvre de la population. On la définit comme une fonction de p $(0 \le p \le 1)$ qui représente simplement le revenu cumulatif en fonction de la part de la population. En tant que paramètre, nous l'exprimons sous la forme

$$L(p) = \frac{1}{\mu_Y} \int_0^p \xi_q \, dq$$

où μ_Y est la moyenne de la population et où ξ_q est la fonction de quantile. Dans le cas d'une grande population sans double, l'expression ci-dessus est représentée approximativement par

$$L_N(p) \approx \sum\nolimits_U \frac{I\{F_i \le p\}\, Y_i}{\mu_N} \frac{1}{N}$$

et estimée par

$$\hat{L}(p) = \sum\nolimits_s \frac{I\{\hat{F}_{hci} \le p\}\, y_{hci}}{\hat{\mu}} \tilde{w}_{hci}$$

où $\hat{\mu} = \sum_s \tilde{w}_{hci}\, y_{hci}$ et $\hat{F}_{hci} = \hat{F}(y_{hci})$.

Par *part du quantile (PQ)*, nous entendons la proportion du revenu total que partage la population associée à un intervalle de quantile $[\xi_{p_1}, \xi_{p_2})$, soit:

$$Q_N(p_1, p_2) \approx \sum\nolimits_U \frac{I\{ep_1 \le F_i \le p_2\}\, Y_i}{\mu_N} \frac{1}{N} = L_N(p_2) - L_N(p_1).$$

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 47-59
Statistique Canada

Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu – Étude empirique

MILORAD S. KOVAČEVIĆ et WESLEY YUNG[1]

RÉSUMÉ

Les mesurent de l'inégalité et de la polarisation du revenu sont essentielles à l'étude de nombreux dossiers économiques et sociaux. La plupart de ces mesures étant des fonctions non linéaires de la fonction de distribution et(ou) des quantiles, on ne peut exprimer leur variance au moyen d'une formule simple et on doit recourir aux méthodes d'estimation de la variance approximative. Dans le présent article, on résume plusieurs méthodes appliquées à l'estimation de la variance de six mesures particulières de l'inégalité et de la polarisation du revenu et on étudie empiriquement leur performance grâce à une étude en simulation fondée sur l'Enquête canadienne sur les finances des consommateurs. Les résultats indiquent que, pour les mesures étudiées, la méthode bootstrap et celle des équations d'estimation donnent de nettement meilleurs résultats que les autres.

MOTS CLÉS: Indice de Gini; ordonnée de la courbe de Lorenz; proportion de faible revenu; indice de polarisation; part du quantile; estimation de la variance par rééchantillonnage; méthode de linéarisation.

1. INTRODUCTION

Les analyses de la répartition du revenu sont des éléments fondamentaux de l'examen de questions socioéconomiques importantes, dont la grandeur de l'inégalité, la pauvreté ou la grandeur de la classe moyenne. De nombreux articles statistiques et économétriques ont été publiés sur le sujet, particulièrement sur les diverses mesures de l'inégalité du revenu et sur leurs propriétés (Sen (1973); Kakwani (1980); Nygärd et Sandström (1981)). Cependant, les auteurs s'efforcent rarement de produire des données sur la variabilité d'échantillonnage des estimations utilisées pour déterminer la grandeur de l'inégalité ou de la polarisation. Or, on doit s'appuyer sur ce genre d'information pour i) déterminer la précision des estimations obtenues à partir de données d'enquête et ii) faire des inférences statistiques officielles sur la répartition du revenu, particulièrement quand on effectue des comparaisons régionales ou chronologiques.

Les mesures de l'inégalité et de la polarisation du revenu étant des paramètres de population finie qui s'expriment sous forme de fonctions des valeurs ordonnées de la population, on ne peut calculer leur variance au moyen d'une formule simple et on doit recourir aux méthodes d'estimation de la variance approximative. En général, les inférences concernant ces mesures, fondées sur un plan d'échantillonnage complexe, englobent une estimation ponctuelle et des intervalles de confiance. Nous étudions ici l'estimation de la variance de certaines de ces mesures, dont les quantiles, le seuil de faible revenu, la proportion de faible revenu, l'ordonné de la courbe de Lorenz, la parts du quantile, l'indice de Gini et l'indice de polarisation.

Dans tout l'article, nous supposons que la population observée est une population finie invariable, autrement dit que chaque unité de la population est associée à un nombre réel, fixe, mais inconnu, à savoir la valeur du revenu gagné par l'unité. Nous supposons aussi que la population est subdivisée en L strates, et nous représentons par N_h le nombre d'unités primaires d'échantillonnage (UPÉ) dans la h-ième strate. Au premier degré d'échantillonnage, nous tirons $n_h(\geq 2)$ UPÉ de la strate h (de façon indépendante d'une strate à l'autre). Enfin, nous supposons qu'on effectue un sous-échantillonnage des UPÉ échantillonnées pour s'assurer que les estimations des totaux des UPÉ, $\hat{Y}_{hc}, c = 1, ..., n_h$; $h = 1, ..., L$ ne soient pas biaisées. La (hci)-ième unité finale d'échantillonnage est liée à la valeur observée de la variable étudiée, y_{hci}, ainsi qu'au poids d'échantillonnage w_{hci}. Nous représentons par $\sum_s = \sum_h \sum_c \sum_i$ la sommation multiple sur toutes les unités finales de l'échantillon tenant compte de tous les degrés d'échantillonnage.

Après avoir passé en revue les définitions fondamentales des mesures étudiées, nous présentons à la section 2 leur estimation ponctuelle conformément au plan de sondage proposé. À la section 3, qui traite de l'estimation de la variance de ces mesures, nous passons en revue les méthodes existantes et nous en décrivons cinq en détail, à savoir la méthode jackknife, les méthodes du groupement de demi-échantillons compensés et du groupement répété de demi-échantillons compensés, la méthode bootstrap et la méthode de linéarisation fondée sur des équations d'estimation. À la section 4, nous décrivons l'étude en simulation basée sur les données de l'Enquête canadienne sur les finances des consommateurs de 1988. Cette étude empirique a pour objectif de comparer les méthodes d'estimation de la variance d'un certain nombre de mesures de l'inégalité du revenu. Nous présentons, résumons et interprétons divers résultats. Enfin, à la section 5, nous présentons nos conclusions.

[1] Milorad S. Kovačević, méthodologiste principale, Division des méthodes d'enquêtes des ménages, et Wesley Yung, méthodologiste sénior, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario) Canada, K1A 0T6.

KALTON, G., et MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 409-428.

LEPKOWSKI, J., KALTON, G., et KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

LITTLE, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, (Vol. 2), (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects *Biometrika*, 70, 41-55.

ROSENBAUM, P.R., et RUBIN, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

TREMBLAY, V. (1986). Critères pratiques pour la définitions des classes de pondération. *Techniques d'enquête*, 12, 91-103.

UNITED STATES BUREAU OF LABOR STATISTICS (1991). News: Consumer Expenditures in 1990. Publication USDL 91-607, United States Department of Labor, Washington, DC.

UNITED STATES BUREAU OF LABOR STATISTICS (1992). BLS Handbook of Methods. Bulletin 2414, United States Department of Labor, Washington, DC.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YANSANEH, I.S., et ELTINGE, J.L. (1993). Construction of adjustment cells based on surrogate items or estimated response propensities. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 538-543.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

2. Construire k cellules de correction dont les limites sont déterminées par les quantiles $k^{-1}j$ estimés de la population de $\hat{\eta}_i$, $j = 1, 2, ..., k-1$. Calculer l'estimation de la moyenne corrigée résultante, $\hat{\bar{Y}}_k$.
3. Répéter (2) pour plusieurs nombre entiers $k > 1$. À mesure que k augmente, repérer le point où $\hat{\bar{Y}}_k$ devient à peu près constante. Compte tenu des résultats de Rosenbaum et Rubin (1984) et des résultats empiriques exposés ici, les valeurs de k proches de 5 pourraient présenter un intérêt particulier.
4. Utiliser des méthodes diagnostiques simples (par exemple $\hat{B}_h$ et $\bar{d}_h$ décrits à la section 3.2) pour repérer les cellules de corrections obtenues par division en quantiles égaux posant éventuellement des problèmes. Si l'application de la méthode diagnostique indique que certaines cellules sont problématiques, essayer d'effectuer une mise au point supplémentaire de ces cellules. Calculer les estimations de $\bar{Y}$ d'après les ensembles perfectionnés de cellules et comparer les nouvelles estimations aux valeurs de $\hat{\bar{Y}}_k$ obtenue au point 3.
5. Évaluer l'effet global de la correction en comparant les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ aux erreurs-types $se(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ et en calculant les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$.
6. Répéter les étapes (1) à (5), au besoin, pour les cellules de correction fondées sur les $\hat{Y}_i$. Comparer les estimations finales de $\bar{Y}$ obtenues par les méthodes fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$.

5.2 Domaines dans lesquels il faut poursuivre les travaux

Les résultats de la présente étude donnent à penser qu'il pourrait être utile de poursuivre les travaux dans deux domaines. Premièrement, le problème que pose la non-réponse aux questions sur le revenu dans le cas de la CE est similaire à celui que pose la non-réponse dans le cas de plusieurs autres enquêtes à grande échelle. Cependant, comme pour toute étude de cas, on devrait éviter de surgénéraliser les résultats empiriques présentés ici. Il serait utile d'appliquer les méthodes diagnostiques que nous décrivons aux problèmes que posent d'autres estimations (par exemple, moyennes croisées) ou à des ensembles de données sur la non-réponse présentant des caractéristiques légèrement différentes (par exemple, taille effective de l'échantillon plus grande ou plus petite, ou distribution plus étalée ou plus serrée des estimations des $\hat{\eta}_i$). Cet exercice apporterait des éclaircissements sur les caractéristiques opératoires des méthodes de construction de cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ dans la pratique. Deuxièmement, il serait intéressant d'étendre l'étude à des problèmes à plusieurs variables (par exemple, rapport entre les données sur le revenu de la deuxième et de la cinquième interview de la CE).

REMERCIEMENTS

Les auteurs remercient Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott et Stephanie Shipp pour les nombreuses discussions fructueuses sur la Consumer Expenditure Survey, et Wayne Fuller, Steve Miller, Geoff Paulin, Stuart Scott, trois arbitres, et le rédacteur, pour leurs commentaires précieux sur des versions antérieures de l'article. Les présents travaux ont été effectués pendant un séjour des auteurs au Bureau of Labor Statistics organisé dans le cadre du ASA/NSF/BLS Research Fellow Program et financé grâce à une bourse de la National Science Foundation (SES-9022443). Les travaux de recherche de Eltinge ont également été financés en partie par une bourse du National Institutes of Health (CA 57030-04). Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas nécessairement les politiques du Bureau of Labor Statistics.

BIBLIOGRAPHIE

CASSEL, C.-M., SÄRNDAL, C.-E., et WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. Dans *Incomplete Data in Sample Surveys*, (Vol. 3), (Éds. W.G. Madow, I. Olkin, et D. Rubin). New York: Academic Press, 143-160.

COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 205-213.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R.J.A., SAMUHEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 168-173.

DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

EZZATI, T., et KHARE, M. (1992). Nonresponse adjustments in a national health survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 339-344.

GARNER, T.I., et BLANCIFORTI, L.A. (1994). Household income reporting: An analysis of U.S. Consumer Expenditure Survey data. *Journal of Official Statistics* 10, 69-91.

GOKSEL, H., JUDKINS, D.R., et MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

Conséquemment, $\hat{\gamma}_k$ reflète la perte d'efficacité résultant de l'utilisation de l'estimateur biaisé, non corrigé $\hat{Y}_1$ au lieu de l'estimateur corrigé, non biaisé $\hat{\bar{Y}}_k$. Cependant, cette interprétation doit être considérée avec prudence, puisqu'elle dépend de la supposition que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$, et, puisque les $\hat{\gamma}_k$ sont des fonctions des termes aléatoires $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$, $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$.

Comme l'a suggéré un arbitre, on pourrait aussi considérer le ratio des erreurs quadratiques moyennes

$$\{\hat{V}(\hat{\bar{Y}}_\eta)\}^{-1}[\hat{V}(\hat{\bar{Y}}_k) + \max\{0, (\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)^2 - \hat{V}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_\eta)\}]$$

où $\hat{\bar{Y}}_\eta$ est égal à l'expression (1.1) avec λ_i remplacé par $(\hat{\eta}_i)^{-1}\lambda_i$. Cette approche équivaudrait à comparer chaque estimation $\hat{\bar{Y}}_k$ fondée sur les cellules à $\hat{\bar{Y}}_\eta$. Cette démarche est appropriée si $\hat{\bar{Y}}_\eta$ est approximativement non biaisé, mais l'absence de biais peut être problématique dans certains cas; à cet égard, consulter Little (1986, p. 146).

La dernière colonne du tableau 1 donne les ratios estimés $\hat{\gamma}_k$ pour des valeurs particulières de k. Pour $k \geq 5$, chaque $\hat{\gamma}_k$ présenté est plus grand que 1.5. Enfin, soulignons que chaque estimation corrigée $\hat{\bar{Y}}_k$ est inférieure à l'estimation non corrigée $\hat{\bar{Y}}_1$. Cette situation tient au fait que, pour une valeur donnée de k, les cellules associées aux probabilités de réponse les plus grandes ont tendance à produire une estimation plus grande de la moyenne $\bar{Y}_{hR}$. Par exemple, pour $k = 5$, les valeurs de $\bar{Y}_{hR}$ sont \$24,333, \$33,729, \$33,398, \$34,620 et \$37,057 pour $h = 1$ (cellule à $\hat{\eta}_i$ la plus faible) à $h = 5$ (cellule à $\hat{\eta}_i$ la plus élevée), respectivement.

4. CELLULES FONDÉES SUR LES VALEURS ESTIMÉES DU REVENU

Les idées diagnostiques générales décrites à la section 3 s'appliquent également aux cellules fondées sur $\hat{Y}_i$. Pour illustrer ce point, nous ajustons séparément des équations de régression pondérée où Y_i = représente le revenu déclaré par les répondants à la deuxième et à la cinquième interviews. Yansaneh et Eltinge (1993) décrivent en détail les calculs, y compris l'estimation des paramètres et des erreurs-types. Nous nous sommes servi des modèles de régression résultants pour calculer les estimations du revenu $\hat{Y}_i$ pour les unités qui ont déclaré complètement et incomplètement leur revenu, puis nous avons regroupé les unités en cellules d'après leur valeur de $\hat{Y}_i$, les limites des cellules étant déterminées par la méthode des quantiles égaux.

Nous présentons au tableau 6 les résultats de l'analyse fondamentale de sensibilité et de la mesure de l'efficacité pour les cellules fondées sur les $\hat{Y}_i$; sa présentation est la même que celle du tableau 1. Les résultats de l'analyse de sensibilité sont qualitativement similaires, mais non identiques, à ceux présentés pour les cellules fondées sur les $\hat{\eta}_i$. Au cours de travaux supplémentaires non décrits en détail ici, nous avons examiné la division de chaque cellule fondée sur les $\hat{Y}_i$ en quantiles égaux. Pour $k \geq 4$, les estimations résultantes de la moyenne et les erreurs-types connexes ne diffèrent pas notablement de celles présentées au tableau 6.

Tableau 6
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles du revenu estimé

Méthode de correction	Estimation ponctuelle	Erreur-type	$ET(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio EQM
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,512	509	106	2.01
$k = 4$ cellules	32,468	512	108	2.14
$k = 5$ cellules	32,473	511	115	2.12
$k = 6$ cellules	32,492	508	117	2.08
$k = 10$ cellules	32,488	510	119	2.07
$k = 15$ cellules	32,478	504	124	2.16
$k = 20$ cellules	32,495	513	124	2.02

Les deux dernières colonnes du tableau 6 permettent de comparer $\hat{\bar{Y}}_k$ à l'estimation non corrigée $\hat{\bar{Y}}_1$. Pour $k \geq 4$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$472, et les erreurs-types estimées, inférieures ou égales à \$124. Les valeurs de la statistique t associée sont toutes supérieures à 3.80. En outre, les ratios des erreurs quadratiques moyennes estimées $\hat{\gamma}_k$ sont tous plus grand que 2.0.

Qui plus est, les cellules fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ produisent des estimations corrigées légèrement différente du revenu moyen, mais les écarts observés ne sont pas statistiquement significatifs pour les seuils de signification α ordinaires. Par exemple, pour $k = 5$, l'écart entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est égal à \$32,630\$ - \$32,473 = \$157, l'erreur-type est égale à \$122 et la statistique t est égale à 1.29. Pareillement, pour $k = 10$, la différence entre les estimations fondées sur les $\hat{\eta}_i$ et sur les $\hat{Y}_i$ est de \$152, avec une erreur-type de \$104. Donc, les données permettent peu de faire la distinction entre les résultats obtenus par les deux méthodes générales de construction de cellules.

Enfin, soulignons qu'un ensemble donné de cellules fondées sur les $\hat{Y}_i$ est fondamentalement lié à une variable Y particulière, comme le revenu de l'unité de consommation. Par conséquent, cet ensemble de cellules ne donnera pas nécessairement de bons résultats pour l'estimation de la moyenne d'une variable Y différente.

5. DISCUSSION

5.1 Résumé des méthodes

Nous examinons dans le présent article certaines méthodes diagnostiques simples permettant de construire des cellules de correction pour la non-réponse. Nous pouvons résumer comme suit la méthodologie.

1. D'après des travaux de modélisation préliminaires et les variables auxiliaires étudiées X_i, calculer la probabilité estimée de réponse $\hat{\eta}_i$ pour chaque unité d'échantillonnage (répondants et non-répondants).

$$(LB_i, UB_i) = ([1 + \exp\{-X_i'\hat{\theta} + 1.96 D_i^{1/2}\}]^{-1}$$
$$[1 + \exp\{-X_i'\hat{\theta} - 1.96 D_i^{1/2}\}]^{-1}),$$

où $\hat{\theta}$ est le vecteur des estimations des paramètres de régression logique, où $D_i = X_i' \hat{V}_\theta X_i$, et où $\hat{V}_\theta$ est la matrice de covariance estimée d'après la pseudorépétition pour $\hat{\theta}$. Représentons par $\bar{d}_h$ la moyenne de l'échantillon pondérée en fonction de λ_i des largeurs des intervalles de confiance $UB_i - LB_i$ pour les unités i de la cellule h, et comparons $\bar{d}_h$ à la largeur de la cellule h. Si la cellule h est relativement large, tant en valeur absolue que comparativement à $\bar{d}_h$, alors, la division de cette cellule peut produire de nouvelles cellules ayant des facteurs de pondération a_h nettement différents. Inversement, si $\bar{d}_h$ est beaucoup plus grand que la largeur de la cellule h, alors, les écarts entre les $\hat{\eta}_i$ dans cette cellule peuvent résulter davantage de l'erreur d'estimation que d'écarts entre les η_i réels. Le cas échéant, une division supplémentaire de la cellule modifiera vraisemblablement peu les facteurs de pondération a_h; et, par conséquent, l'estimateur de $\bar{Y}$ corrigé pour la non-réponse obtenu variera assez peu.

Nous présentons aux tableaux 4 et 5 les limites des cellules, les largeurs des cellules, $\bar{d}_h$, ainsi que les valeurs de a_h pour $k = 5$ et $k = 10$, respectivement. Pour $k = 5$, la largeur des cellules 2 à 5 n'est pas grande comparativement aux valeurs de $\bar{d}_h$. Essentiellement, chacune de ces cellules est divisée en deux pour produire le cas où $k = 10$ cellules. Les paires résultants de a_h pour $k = 10$ sont assez proches des valeurs correspondantes de a_h dans les cellules 2 à 5 pour $k = 5$.

Tableau 4
Limites des cellules fondées sur la probabilité estimée de réponse, largeur moyenne des cellules, largeur des intervalles de confiance et facteur de correction pour la non-réponse, $k = 5$

h	Limite inférieure	Limite Supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.810	0.426	0.197	1.35
2	0.810	0.861	0.051	0.139	1.20
3	0.861	0.894	0.033	0.110	1.13
4	0.894	0.924	0.030	0.088	1.08
5	0.924	0.994	0.070	0.067	1.07

En revanche, pour $k = 5$, la cellule 1 est plus de deux fois plus large que $\bar{d}_1$. Quand $k = 10$, cette cellule est divisée en cellules plus petites ayant des facteurs de pondération corrigés pour la non-réponse a_h légèrement différents, soit 1.45 et 1.27, respectivement. Cependant, les estimations correspondantes de la moyenne cellulaire sont assez proches, à savoir $\bar{Y}_{1R}$ = \$24,045 et $\bar{Y}_{2R}$ = \$24,582 pour $k = 10$. Donc, dans cet exemple, les estimations corrigées pour la non-réponse $\hat{\bar{Y}}_5$ et $\hat{\bar{Y}}_{10}$ sont relativement proches, parce que quatre des cinq divisions cellulaires entraînent une variation assez faible des poids et que la cinquième produit des cellules dont la moyenne est similaire.

Tableau 5
Limites des cellules fondées sur la probabilité estimée de réponse, largeur des cellules, largeur moyenne des intervalles de confiance et facteur de correction pour la non-réponse, $k = 10$

h	Limite inférieure	Limite supérieure	Largeur de la cellule	$\bar{d}_h$	a_h
1	0.384	0.762	0.378	0.220	1.45
2	0.762	0.810	0.048	0.174	1.27
3	0.810	0.840	0.030	0.146	1.21
4	0.840	0.861	0.021	0.132	1.19
5	0.861	0.878	0.017	0.111	1.14
6	0.878	0.894	0.016	0.108	1.11
7	0.894	0.908	0.014	0.093	1.09
8	0.908	0.924	0.016	0.083	1.08
9	0.924	0.944	0.020	0.072	1.08
10	0.944	0.994	0.050	0.062	1.06

Enfin, les facteurs a_h du tableau 5 indiquent que les taux moyens de réponse dans le cas de $k = 10$ cellules se situent dans une fourchette raisonnable, allant de $(1.45)^{-1} = 0.69$ à $(1.06)^{-1} = 0.94$. Certains ensembles de données sur la non-réponse sont caractérisés par une fourchette plus large, donc plus susceptibles de produire des écarts plus prononcés après la division des cellules. Inversement, d'autres sont caractérisés par une distribution plus serrée des probabilités de réponse, donc moins susceptibles d'être affectés notablement par la division des cellules.

3.3 Comparaison des estimations corrigées d'après les cellules aux estimations non corrigées

Pour conclure l'évaluation des cellules fondées sur les $\hat{\eta}_i$, nous comparons les estimations corrigées $\hat{\bar{Y}}_k$ aux estimations non corrigées $\hat{\bar{Y}}_1$. Premièrement, le tableau 1 indique que, pour les valeurs déclarées de $k \geq 5$, les différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont supérieures ou égales à \$303. Deuxièmement, pour $k \geq 5$, les erreurs-types estimées des différences $\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k$ sont toutes inférieures ou égales à \$138, et les valeurs correspondantes de la statistique t sont toutes supérieures à 2.44. Donc, pour $k = 5$ par exemple, un test formel de l'hypothèse H_0: $E(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_5) = 0$ mènerait au rejet de cette dernière pour les seuils de signification types, ce qui signifie que la méthode des cellules de correction produit une modification importante de l'estimation du revenu moyen.

De surcroît, une comparaison grossière de l'efficacité de $\hat{\bar{Y}}_1$ et $\hat{\bar{Y}}_k$ se dégage du ratio des erreurs quadratiques moyennes estimées

$$\hat{\gamma}_k = \{\hat{V}(\hat{\bar{Y}}_k)\}^{-1}[\hat{V}(\hat{\bar{Y}}_1) + \max\{0, (\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)^2 - \hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)\}]$$

où $\hat{V}(\hat{\bar{Y}}_1)$, $\hat{V}(\hat{\bar{Y}}_k)$, et où $\hat{V}(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_k)$ sont les estimations de la variance basée sur la pseudorépétition pour les moyennes indiquées. Pour interpréter ce ratio, supposons pour le moment que $\hat{\bar{Y}}_k$ est un estimateur approximativement non biaisé de $\bar{Y}$. Alors, $\hat{\gamma}_k$ est un estimateur de l'erreur quadratique moyenne de l'estimateur non corrigé $\hat{\bar{Y}}_1$, comparativement à l'erreur quadratique moyenne de $\hat{\bar{Y}}_k$.

diagnostique proposée mène au dépistage des «cellules problématiques» éventuelles et à la construction d'un ensemble plus perfectionné de cellules de correction que nous appelons C_2. La comparaison des estimations de $\bar{Y}$ fondées sur C_1 et C_2 permet alors de décider quel est l'ensemble de cellules de correction fondées sur $\hat{\eta}_i$ qui convient le mieux.

3.2.1 Évaluation du biais à l'intérieur des cellules

Comme nous l'avons fait remarquer à la section 1.2, un estimateur corrigé $\hat{\bar{Y}}_k$ donné réduit, mais n'élimine pas complètement, le biais dû à la non-réponse et le biais résiduel de $\hat{\bar{Y}}_k$ dépend du biais qui entache les estimations de la moyenne intracellulaire $\bar{Y}_{hR}$. Considérons l'autre estimateur de la moyenne intracellulaire

$$\bar{Y}_{h\eta} = \left(\sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i \right)^{-1} \sum_{i \in s_h} \hat{\eta}_i^{-1} \lambda_i R_i Y_i. \qquad (3.1)$$

Si les estimations $\hat{\eta}_i$ étaient égales aux probabilités réelles de réponse η_i, alors (3.1) serait un estimateur approximativement non biaisé de la moyenne réelle de la sous-population $\bar{Y}_h$. Le cas échéant, un estimateur du biais intracellulaire $\mathrm{E}(\bar{Y}_{hR} - \bar{Y}_h)$ serait $\hat{B}_h = \bar{Y}_{hR} - \bar{Y}_{h\eta}$, et l'estimateur correspondant du biais global $\mathrm{E}(\hat{\bar{Y}}_k - \bar{Y})$ serait $\hat{B} = (\sum_{h=1}^{k} \sum_{j \in s_h} \lambda_j)^{-1} \sum_{h=1}^{k} (\sum_{j \in s_h} \lambda_j) \hat{B}_h$.

Puisque les valeurs de $\hat{\eta}_i$ sont sujettes à des erreurs d'estimation, les termes $\hat{B}_h$ et $\hat{B}$ ne donnent qu'une indication partielle des problèmes éventuels de biais. Par exemple, une grande valeur de $\hat{B}_h$ pourrait être le reflet d'un biais important entachant $\hat{\bar{Y}}_{hR}$ ou de biais entachant l'autre estimateur $\bar{Y}_{h\eta}$, à cause des erreurs $\hat{\eta}_i - \eta_i$; lire les commentaires de mise en garde de Little (1986, p. 146) concernant l'utilisation directe des poids $\hat{\eta}_i^{-1}$ pour établir l'estimation corrigée de $\bar{Y}$. Donc, si on observe une grande valeur de $\hat{B}_h$, cela vaut la peine d'envisager la mise au point de la cellule h, mais la décision finale quant à l'utilisation de l'ensemble perfectionné de cellules ainsi obtenu dépendra du fait que cet ensemble produit ou non une estimation nettement différente de la moyenne globale $\bar{Y}$.

Nous présentons aux tableaux 2 et 3 les valeurs de $\hat{B}_h$, les erreurs-types associées et les valeurs de la statistique t pour les cellules formées par division en quantiles égaux pour $k = 3$ et $k = 5$, respectivement. Soulignons que, dans le cas où $k = 3$, le test diagnostic s'appliquant à $\hat{B}_h$ indique une contribution éventuelle au biais pour la cellule de rang le plus bas. Cette observation concorde avec l'idée énoncée à la section 3.1 selon laquelle $k = 3$ cellules pourrait ne pas donner une correction satisfaisante pour la non-réponse. En outre, la valeur correspondante de $\hat{B}$ est 111, avec une erreur-type de 75; cette valeur de $\hat{B}$ est très proche de la différence $\hat{\bar{Y}}_3 - \hat{\bar{Y}}_5 = 106$ des estimations $\hat{\bar{Y}}_3$ et $\hat{\bar{Y}}_5$ du tableau 1.

À la lumière des résultats qui précèdent, nous avons divisé en deux la cellule à faible $\hat{\eta}_i$ du cas où $k = 3$. Nous avons déterminé les limites supérieures des deux nouvelles cellules (soit, $h = 1'$ et $h = 1''$) grâce aux quantiles estimés 0.167 et 0.333 de la population de $\hat{\eta}_i$. Les valeurs résultantes de $\hat{B}_h$ et les erreurs-types sont 90 et 197 pour la cellule $1'$, et -42 et 79, pour la cellule $1''$. En outre, l'ensemble ainsi perfectionné de quatre cellules donne $\hat{B} = 30$, avec une erreur-type de 75, et l'estimation corrigée de $\bar{Y}$, égale à \$32,652, ainsi que l'erreur-type, égale à \$518, sont proches des valeurs obtenues par la méthode de division en quantiles égaux pour $k = 5$.

Tableau 2
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 3$

h	$\hat{B}_h$	$\mathrm{et}(\hat{B}_h)$	$t = \hat{B}_h/\mathrm{et}(\hat{B}_h)$
1	269	136	1.98
2	-19	43	0.44
3	84	45	1.87

Tableau 3
Statistiques $\hat{B}_h$ cellulaires pour les cellules fondées sur les probabilités, $k = 5$

h	$\hat{B}_h$	$\mathrm{et}(\hat{B}_h)$	$t = \hat{B}_h/\mathrm{et}(\hat{B}_h)$
1	96	217	0.44
2	-72	116	-0.62
3	-52	56	-0.93
4	-16	27	-0.59
5	98	50	1.96

Contrairement aux résultats obtenus pour $k = 3$, les valeurs de $\hat{B}_h$ obtenues pour $k = 5$ posent assez peu de problèmes, sauf, éventuellement, dans le cas de la cellule $h = 5$, pour laquelle la valeur de la statistique t est égale à 1.96. Pour $k = 5$, la valeur de $\hat{B}$ est 11, avec une erreur-type de 93. Une division supplémentaire de la cellule $h = 5$ n'a pas modifié notablement l'estimation de $\bar{Y}$ ou de l'erreur-type associée. Les valeurs de $\hat{B}_h$ résultant de la division des cellules correspondant à une valeur de k plus grande par la méthode des quantiles égaux présentent encore moins de signes de l'existence d'un biais intracellulaire. Par exemple, pour $k = 6$, la statistique t est inférieure ou égale à 1.65 pour les valeurs de $\hat{B}_h$ de chacune des six cellules et, pour $k = 10$, la statistique t est inférieure ou égale à 1.54 pour les valeurs de $\hat{B}_h$ de chacune des dix cellules.

3.2.2 Comparaison entre la largeur des cellules et la précision des estimations η_i

La comparaison de la largeur des cellules de correction à la largeur des intervalles de confiance associés aux probabilités de réponse η_i fournit une autre méthode diagnostique pour repérer les cellules problématiques éventuelles. Premièrement, représentons par $a_h = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i$, le facteur de correction pour la non-réponse appliqué aux unités répondantes de la cellule h. Deuxièmement, conformément aux résultats types de la régression logistique, notons qu'un intervalle de confiance d'environ 95% pour η_i est

des données de la CE. Conformément à cette approche, nous ne ferons la distinction entre les données de la deuxième et de la cinquième interviews dans le présent article que pour construire les modèles de $\hat{\eta}_i$ et $\hat{Y}_i$.

Ici, nous avons utilisé les données des rapports de la deuxième et de la cinquième interviews pour toutes les unités de consommation pour lesquelles une deuxième interview était prévue en 1990. Les données de la deuxième interview se rapportent à 5,125 unités et celles de la cinquième, à 5,093 unités. Pour chaque unité interviewée (ayant déclaré complètement ou incomplètement son revenu), nous avons tiré des enregistrements du BLS des données sur un grand nombre de variables démographiques et de variables de dépenses que nous avons utilisées comme variables auxiliaires dans les travaux de modélisation décrits aux sections 3 et 4 ci-après. À la deuxième ainsi qu'à la cinquième interview, environ 14% des unités de consommation interviewées ont déclaré incomplètement leur revenu.

3. CELLULES FONDÉES SUR LES PROBABILITÉS ESTIMÉES DE RÉPONSE

Nous examinons pour commencer la construction de cellules de correction fondées sur les probabilités estimées de réponse. Nous avons ajusté séparément les modèles de régression logistique utilisés pour calculer la probabilité qu'une unité déclare complètement son revenu $\eta_i = \eta(X_i)$ aux données de la deuxième et de la cinquième interviews décrites à la section 2. Les détails de l'ajustement des modèles, y compris l'estimation des paramètres et le calcul des erreurs-types, sont décrits dans Yansaneh et Eltinge (1993). Nous avons calculé toutes les estimations de la variance par la méthode de pseudorépétition décrite à la section 2. Nous nous sommes servis des modèles résultant des ajustements finals pour estimer, pour chaque unité ayant participé à la deuxième et à la cinquième interviews, les probabilités de déclarer complètement le revenu $\hat{\eta}_i$. Conformément à la stratégie décrite à la section 1.3, nous avons regroupé les unités selon la valeur de $\hat{\eta}_i$ en k cellules dont nous avons défini les limites par la méthode des quantiles égaux.

3.1 Analyse initiale de la sensibilité au nombre choisi de cellules

Les trois premières colonnes du tableau 1 donnent les estimations ponctuelles corrigées $\hat{\bar{Y}}_k$ du revenu moyen et les erreurs-types associées pour plusieurs valeurs de k. La comparaison de ces estimations ponctuelles indique dans quelle mesure les estimations corrigées sont sensibles au choix d'une valeur particulière de k. Pour $k \geq 5$, les estimations ponctuelles présentées sont relativement stables, variant de \$32,630 à \$32,664. Cette observation concorde avec l'idée énoncée à la section 1.3 selon laquelle $k = 5$ cellules pourrait fournir la plupart de la réduction effective du biais produite par une méthode donnée de construction des cellules; consulter Rosenbaum et Rubin (1984, section 1 et appendice A) pour certains renseignements mathématiques connexes.

En outre, soulignons que pour $k \geq 3$, l'erreur-type qui entache $\hat{\bar{Y}}_k$ est également assez stable, variant de \$508 à \$530. Cette observation contredit en partie l'idée générale selon laquelle le choix du nombre approprié de cellules s'appuie sur un compromis entre le biais et la variance. En ce qui concerne l'ensemble de données examinées ici, il semble que la réduction effective du biais se produise assez rapidement (disons, pour $k = 5$,), alors qu'une augmentation considérable de la variance ne survient qu'après qu'on ait dépassé la valeur $k = 20$. Ce résultat n'est pas irréaliste, puisque, même pour $k = 20$, le nombre de réponses sur le revenu par cellule demeure assez grand (variant de 461 à 569), donc ne donne pas lieu au problème général de l'estimateur instable associé à un nombre croissant de cellules peu peuplées. Par contre, le compromis entre le biais et la variance pourrait poser des problèmes plus graves pour des valeurs assez faibles de k dans le cas d'applications où la taille effective de l'échantillon est plus petite, comme l'estimation de petites sous-populations.

Tableau 1
Estimations corrigées du revenu moyen quand les limites des cellules sont déterminées d'après les quantiles des probabilités estimées de réponse

Nombre de cellules	Estimation ponctuelle	Erreur-type	$\mathrm{ET}(\hat{\bar{Y}}_k - \hat{\bar{Y}}_1)$	Ratio $\mathrm{EQM}(\hat{\gamma}_k)$
Non corrigée ($k = 1$)	32,967	569	S/O	S/O
$k = 3$ cellules	32,736	530	112	1.30
$k = 4$ cellules	32,779	518	122	1.28
$k = 5$ cellules	32,630	523	138	1.53
$k = 6$ cellules	32,664	515	122	1.51
$k = 10$ cellules	32,640	514	116	1.58
$k = 15$ cellules	32,638	515	118	1.58
$k = 20$ cellules	32,634	508	118	1.63

3.2 Deux méthodes diagnostiques simples applicables aux cellules

Pour compléter l'analyse de sensibilité qui précède, il est utile d'examiner certains ensembles de cellules de correction plus en détail. Représentons par $C_1 = \{s_1, ..., s_k\}$ un ensemble donné de cellules de correction à examiner, comme les cellules créées par la division en quantiles égaux pour $k = 3$ ou $k = 5$ décrites à la section 3.1. Nous pouvons perfectionner les cellules contenues dans l'ensemble C_1 en effectuant une division en quantiles égaux pour une valeur plus grande de k ou en divisant directement une ou plusieurs cellules de l'ensemble C_1. Ce perfectionnement pourrait être utile quand les observations empiriques indiquent que 1) l'estimateur de la moyenne de la cellule $\bar{Y}_{hR}$ risque d'être fortement biaisé ou que 2) une cellule est large comparativement à la précision avec laquelle les valeurs η_i sont estimées. Nous décrivons aux sous-sections 3.2.1 et 3.2.2 deux méthodes diagnostiques simples qui permettent de résoudre les problèmes (1) et (2), respectivement. Dans chaque sous-section, la méthode

permet de remplacer une valeur manquante dans une cellule de correction donnée en sélectionnant au hasard des répondants repères dans la même cellule. Parallèlement à (1.1) et à (1.2), l'estimateur résultant de la moyenne est $\hat{\bar{Y}}_{\text{imp}} = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s} \lambda_i Y_i^*$, où Y_i^* représente une valeur observée ou imputée, selon le cas. En pratique, on s'appuie souvent sur la correction par pondération pour tenir compte de la non-réponse des unités et sur l'imputation, pour tenir compte de la non-réponse à une question. Cependant, pour un ensemble donné de cellules, l'estimateur ponctuel corrigé par pondération (1.2) et l'estimateur par imputation $\hat{\bar{Y}}_{\text{imp}}$ sont entachés du même biais approximatif (1.3). Par souci de simplicité, la suite du présent article portera avant tout sur la correction par pondération, mais il ne faut pas perdre de vue que, pour un ensemble donné de cellules, le problème de réduction du biais est le même qu'on se serve de ces cellules pour effectuer la correction par pondération ou par simple imputation «hot-deck».

1.4 Plan de l'article

Nous examinons dans le présent article certains détails de l'application des méthodes de construction de cellules fondées sur la probabilité estimée de réponse et sur la réponse estimée. Nous accordons une attention particulière aux méthodes diagnostiques permettant de déceler les problèmes que pose un ensemble particulier de cellules, et nous justifions et illustrons ces méthodes en décrivant en détail leur application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. À la section 2, nous donnons certains renseignements généraux sur le problème de la non-réponse concernant le revenu. À la section 3, nous décrivons et appliquons plusieurs méthodes diagnostiques, y compris la comparaison des estimations $\hat{\bar{Y}}_k$ et des erreurs-types pour plusieurs valeurs de k (section 3.1), l'évaluation partielle du biais intracellulaire (section 3.2.1), l'évaluation de la largeur des cellules comparativement à la précision des estimations $\hat{\eta}_i$ (section 3.2.2) et la comparaison des estimations corrigées et non corrigées de la moyenne $\hat{\bar{Y}}_k$ et $\hat{\bar{Y}}_1$ (section 3.3). À la section 4, nous montrons qu'on peut appliquer des méthodes diagnostiques semblables à la correction des cellules fondées sur les chiffres prévus de revenu $\hat{Y}_i$, et nous comparons les estimations du revenu moyen calculé d'après les cellules fondées sur les probabilités estimées, d'une part, et sur le revenu estimé, d'autre part. À la section 5, nous résumons les idées principales qui sous-tendent le présent article et mentionnons certains domaines dans lesquels il conviendrait de poursuivre les travaux.

2. NON-RÉPONSE CONCERNANT LE REVENU DANS LE CAS DE LA U.S. CONSUMER EXPENDITURE SURVEY

2.1 Consumer Expenditure Survey, méthodes de pondération et estimation de la variance

La U.S. Consumer Expenditure Survey (CE) est une enquête à plan d'échantillonnage stratifié à plusieurs degrés avec renouvellement effectuée par le Census Bureau pour le Bureau of Labor Statistics (BLS). Les éléments de l'échantillon sont des «unités de consommation», grossièrement équivalentes aux ménages. Durant l'enquête, on demande à chaque unité d'échantillonnage sélectionnée de participer à cinq interviews. La méthode actuelle de pondération de la CE tient compte des probabilités de sélection initiale, de la correction pour la non-interview, de la stratification a posteriori fondée sur plusieurs variables démographiques et de mises au point supplémentaires; consulter Zieschang (1990) et le United States Bureau of Labor Statistics (1992). En raison de la complexité des travaux de pondération de la CE, le BLS a décidé d'utiliser des estimateurs de la variance fondés sur des méthodes de pseudorépétition à 44 échantillons répétés. Cette pseudorépétition est approximativement équivalente à la répétition compensée type (Wolter 1985, chapitre 3). Toutes les erreurs-types mentionnées ici sont fondées sur la méthode de pseudorépétition, toutes les étapes supplémentaires d'estimation des paramètres et de correction de la pondération étant exécutées séparément pour chaque répétition.

2.2 Non-réponse concernant le revenu

En général, on estime que la correction pour la non-interview incluse dans la méthode de pondération de la CE tient compte comme il convient de la non-réponse d'une unité, comme l'absence de contact ou le refus de participer à une interview particulière. Donc, nous ne nous pencherons plus sur la question de la non-réponse d'une unité ici. Cependant, le BLS craint que les estimations du revenu moyen soient entachées d'un biais dû à la non-réponse partielle aux questions sur le revenu de la CE. Suivent certains renseignements généraux.

Les données détaillées sur le revenu sont collectées durant les deuxième et cinquième interviews de la CE et sont utilisées pour produire des estimations du revenu moyen des unités de consommation (U.S. Bureau of Labor Statistics, 1991) et d'autres paramètres. Les données sur le revenu de la CE sont collectées grâce à un ensemble complexe de questions et le taux de non-réponse à ces questions est relativement élevé. Pour donner une indication sommaire de la réponse ou de la non-réponse à l'ensemble complet de questions sur le revenu, le BLS classifie chaque unité de consommation qui participe à la seconde ou à la cinquième interview comme déclarant complètement ou incomplètement leur revenu. La définition officielle de l'«unité déclarant complètement son revenu» est relativement compliquée; pour une discussion détaillée, consulter Garner et Blanciforti (1994). La méthode appliquée à l'heure actuelle par le BLS pour estimer le revenu moyen consiste à utiliser la réponse moyenne non corrigée $\hat{\bar{Y}}_1$ définie par (1.1), où les R_i représentent les indicateurs de déclaration complète du revenu, Y_i représente le revenu et les poids λ_i correspondent à ceux décrits à la section 2.1. La moyenne pondérée $\hat{\bar{Y}}_1$ est calculée d'après les données de la deuxième ainsi que de la cinquième interview pour une période de référence précise, mais ne s'appuie pas directement sur la structure par panel

en k «cellules de correction» U_h, et l'échantillon s, en groupes correspondants s_h, puis on utilise l'estimateur corrigé

$$\hat{\bar{Y}}_k \stackrel{\text{def}}{=} \sum_{h=1}^{k} w_h \bar{Y}_{hR}, \tag{1.2}$$

où $w_h = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s_h} \lambda_i$ et $\bar{Y}_{hR} = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i R_i Y_i$. Soulignons que, si $k = 1$, alors les estimateurs (1.1) et (1.2) sont identiques. Pour une discussion générale des méthodes basées sur les cellules de correction, consulter, par exemple, Cassel, Särndal et Wretman (1983), Oh et Scheuren (1983), et Kalton et Maligalig (1991).

L'estimateur corrigé $\hat{\bar{Y}}_k$ est entaché d'un biais résiduel dû à la non-réponse approximativement égal à

$$N^{-1} \sum_{h=1}^{k} \bar{\eta}_h^{-1} \sum_{i \in U_h} (\eta_i - \bar{\eta}_h)(Y_i - \bar{Y}_h), \tag{1.3}$$

où N_h représente le nombre d'unités dans U_h et où $(\bar{\eta}_h, \bar{Y}_h) = N_h^{-1} \sum_{i \in U_h} (\eta_i, Y_i)$. Par conséquent, on préfère construire des cellules telles que la covariance entre η_i et Y_i est approximativement nulle dans chaque cellule. En pratique, on s'efforce de le faire en construisant des cellules qui sont approximativement homogènes en regard des probabilités de réponse η_i ou des réponses aux questions Y_i, ou des deux. Dans certains cas, on définit *a priori* des ensembles «naturels» de cellules grâce à des combinaisons de variables de classification connues tant pour les répondants que pour les non-répondants. Par exemple, Ezzati et Khare (1992) utilisent 72 cellules définies selon l'âge, la race, la religion, la situation d'urbanisation et la taille du ménage pour apporter des corrections pour la non-réponse à une partie des données de la National Health and Nutrition Examination Survey. Toutefois, fort souvent, en pratique, la liste des variables susceptibles d'être utilisées pour construire les cellules est assez longue, ce qui peut produire un nombre considérable de cellules contenant peu de répondants, voire aucun. Par conséquent, plusieurs auteurs ont mis au point des méthodes permettant de déterminer les variables de classification les moins importantes et à regrouper les cellules de correction peu peuplées de façon à ce que chaque cellule retenue soit raisonnablement homogène. Consulter, par exemple, Tremblay (1986), Lepkowski, Kalton et Kasprzyk (1989), Kalton et Maligalig (1991), Goskel, Judkins et Mosher (1991) et la discussion connexe sur le regroupement des strates a posteriori de Little (1993). De surcroît, les méthodes axées sur la création de cellules de correction sont apparentées à d'autres, comme les méthodes de correction fondées sur la régression [consulter Rao (1996, section 2.4) et les auteurs qu'il cite] et la méthode itérative généralisée (Deville, Särndal et Sautory 1993).

1.3 Cellules de correction fondées sur l'estimation de la propension à répondre ou sur les réponses prévues

Comme, en principe, les cellules de correction sont approximativement homogènes, on pourrait argumenter que de telles cellules définissent implicitement un modèle pour les valeurs de η_i ou de Y_i, ou des deux. Une modélisation plus explicite mène à deux méthodes connexes de création des cellules. En premier lieu, représentons par X_i un vecteur de variables auxiliaires observées pour les unités d'échantillonnage i répondantes ainsi que non répondantes et servons-nous des valeurs observées dans l'échantillon (R_i, X_i) pour ajuster un modèle pour $\eta_i = \eta(X_i)$ par régression linéaire, logistique ou probit. Puis, construisons les cellules de l'échantillon s_h en groupant les unités d'échantillonnage selon la probabilité estimée de réponse $\hat{\eta}_i$. Comme deuxième solution, considérons la régression des réponses Y_i en fonction d'un vecteur auxiliaire X_i pour produire les réponses estimées $\hat{Y}_i$ pour les unités d'échantillonnage répondantes ainsi que non répondantes. Puis, construisons les cellules de l'échantillon s_h en groupant les unités selon les valeurs de $\hat{Y}_i$.

Ces deux méthodes ont été proposées par Little (1986) comme extension des travaux de Rosenbaum et de Rubin (1983, 1984) sur les scores de propension calculés d'après des données d'observation. Consulter aussi David, Little, Samuhel et Triest (1983). Au départ, les méthodes ont été élaborées dans le contexte d'un modèle, mais elles s'étendent directement au cadre de référence actuel. Selon Little (1986), l'utilisation de cellules fondées soit sur les valeurs de $\hat{\eta}_i$, soit sur celles de $\hat{Y}_i$, pourrait réduire le biais dû à la non-réponse et celle des cellules fondées sur $\hat{Y}_i$ permettrait aussi de contrôler la variance. En outre, dans certains cas, la méthode des cellules fondées sur $\hat{\eta}_i$ et sur $\hat{Y}_i$ est plus souple que celle des cellules définies *a priori*. De surcroît, les cellules de correction fondées sur $\hat{Y}_i$ sont reliées conceptuellement aux notions de stratification optimale (consulter, par exemple, Cochran 1977, sections 5A.7 et 5A.8).

Little (1986) ne propose pas de règle particulière pour déterminer la construction des cellules. Cependant, en s'inspirant des travaux connexes sur les données d'observation effectuées par Cochran (1968) et par Rosenbaum et Rubin (1984), on pourrait envisager de répartir les unités en cellules définies en se fondant sur les quantiles $k^{-1}j$ estimés des populations de $\hat{\eta}_i$ ou de $\hat{Y}_i$ où $j = 1, 2, ..., k-1$. Cette méthode des quantiles égaux permet de contrôler dans une certaine mesure le nombre prévu de répondants dans chaque cellule. En outre, la lecture des deux références susmentionnées donne à penser que, pour un ensemble donné de prédicteurs X_i, on peut réaliser la plupart de la réduction faisable du biais grâce à un nombre relativement faible de cellules, disons $k = 5$. Une étude de cas effectuée par Czajka, Hirabayashi, Little et Rubin (1992) comprend la construction de $k = 6$ cellules de correction fondées sur les $\hat{\eta}_i$ dans chacune de plusieurs strates en appliquant des règles un peu plus complexes que la règle des quantiles égaux considérée ici. Cependant, il ne faut pas surinterpréter le fait qu'un petit nombre de cellules pourrait être adéquat. Par exemple, si on omettait un régresseur important, les estimateurs corrigés d'après les cellules pourraient être entachés d'un biais résiduel important, quel que soit le nombre utilisé de cellules fondées sur la probabilité de réponse ou sur la réponse estimée.

Enfin, on peut remplacer la correction par pondération par l'imputation. Par exemple, la simple imputation «hot-deck»

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 37-45
Statistique Canada

Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey

JOHN L. ELTINGE et IBRAHIM S. YANSANEH[1]

RÉSUMÉ

Les auteurs décrivent certaines méthodes diagnostiques simples utilisées pour guider la construction de cellules de correction pour la non-réponse. S'inspirant des travaux de Little (1986), ils étudient la construction de cellules de correction par regroupement d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la réponse estimée aux questions de l'enquête. Ils examinent plus particulièrement l'évaluation de la sensibilité des estimations corrigées de la moyenne à la variation de k, c'est-à-dire le nombre de cellules utilisées, le dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, la comparaison des estimations corrigées et non corrigées de la moyenne et la comparaison des estimations obtenues au moyen des cellules fondées sur la probabilité estimée de réponse, d'une part, et sur la réponse estimée aux questions, d'autre part. Les auteurs justifient les méthodes proposées et les illustrent par une application à l'estimation du revenu moyen des unités de la U.S. Consumer Expenditure Survey.

MOTS CLÉS: Données incomplètes; données manquantes; quasi-randomisation; propension à répondre; analyse de sensibilité; correction par pondération.

1. INTRODUCTION

1.1 Énoncé du problème

Les analystes d'enquête recourent souvent à la construction de cellules de correction pour tenir compte de la non-réponse. L'idée générale consiste à définir d'abord des groupes, ou «cellules», d'unités d'échantillonnage qu'on estime présenter à peu près la même probabilité de réponse ou produire à peu près la même valeur pour une question particulière, comme celle sur le revenu, puis, à corriger par pondération ou à effectuer une simple imputation «hot-deck» dans chaque cellule de correction. L'estimateur corrigé obtenu d'une moyenne ou d'un total de la population est alors entaché d'un biais dû à la non-réponse approximativement nul, à condition que les covariances intracellulaires entre les réponses aux questions et les probabilités de réponse soient approximativement nulles.

Certains travaux antérieurs sur la correction pour la non-réponse consistaient à créer des cellules de correction en groupant des variables de classification démographiques ou géographiques simples. Cependant, Little (1986) et d'autres chercheurs ont étudié la construction de cellules par groupement direct d'unités d'échantillonnage selon la probabilité estimée de réponse ou selon la valeur estimée des réponses. Dans le présent article, nous examinons certaines méthodes diagnostiques simples qui facilitent l'application de ces idées à la création de cellules. Nous nous concentrons surtout sur la sensibilité des résultats au nombre de cellules utilisées, au dépistage de cellules particulières qui nécessitent une mise au point supplémentaire, à la comparaison des estimations corrigées et non corrigées de la moyenne et à la comparaison des estimations obtenues d'après les cellules fondées sur les probabilités estimées, d'une part, et sur les réponses estimées aux questions, d'autre part. Nous illustrons ces méthodes diagnostiques grâce aux données sur le revenu collectées dans le cadre de la U.S. Consumer Expenditure Survey.

1.2 Notation, biais dû à la non-réponse et cellules de correction

Représentons par U une population donnée de taille N et les questions d'enquête par Y_i, $i \in U$; et considérons l'estimation de la moyenne de la population $\bar{Y} = N^{-1}\sum_{i \in U} Y_i$. Tirons un échantillon s de taille n de la population U et représentons par π_i la probabilité que l'unité i soit incluse dans l'échantillon.

Supposons que la non-réponse satisfait le modèle de quasi-randomisation qui suit (Oh et Scheuren 1983). Posons que R_i est une variable indicatrice égale à 1 si l'unité d'échantillonnage i choisie est un répondant et égale à 0, autrement. Enfin, supposons que les R_i sont des variables aléatoires de Bernoulli (η_i) mutuellement indépendantes, pour lesquelles les probabilités fixes de réponse η_i peuvent varier d'une unité à l'autre. En outre, définissons les poids de sondage $\lambda_i = \pi_i^{-1}$ et la réponse moyenne non corrigée pondérée selon le plan de sondage

$$\hat{\bar{Y}}_1 \stackrel{\text{def}}{=} \left(\sum_{i \in s} \lambda_i R_i\right)^{-1} \sum_{i \in s} \lambda_i R_i Y_i. \tag{1.1}$$

À cause d'écarts entre les η_i, l'estimateur non corrigé $\hat{\bar{Y}}_1$ est entaché d'un biais dû à la non-réponse approximativement égal à $N^{-1}\bar{\eta}^{-1}\sum_{i \in U}\eta_i(Y_i - \bar{Y})$, où $\bar{\eta} = N^{-1}\sum_{i \in U}\eta_i$ et où les espérances mathématiques sont déterminées à la fois sur le plan de sondage original et sur le modèle de quasi-randomisation. Pour réduire ce biais, on partage souvent la population

[1] John L. Eltinge, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.; Ibrahim S. Yansaneh, Westat, 1650 Research Blvd., Rockville, MD 20850-3195, U.S.A.

REMERCIEMENTS

Pedro L.D. Nascimento Silva exprime sa gratitude à CVCP-UK, CNPq-Brasil et IBGE-Brasil pour leur appui financier. Les auteurs remercient Ray Chambers, Danny Pfeffermann, Jon Rao, Michael Bankier et deux arbitres anonymes pour leurs commentaires. Michael Bankier a également offert un appui précieux en fournissant la documentation et le logiciel concernant sa méthode GLSEP.

BIBLIOGRAPHIE

BANKIER, M.D. (1990). Two Step Generalized Least Squares Estimation. Ottawa: Statistique Canada, Division des méthodes d'enquêtes sociales, rapport interne.

BANKIER, M.D., RATHWELL, S., et MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Ottawa: Statistique Canada, document de travail, DMES, 92-007E.

BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

COCHRAN, W.G. (1977). *Sampling Techniques* (3ième éd.). New York: John Wiley & Sons.

DENG, L.Y., et WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

DEVILLE, J.C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DUNSTAN, R., et CHAMBERS, R.L. (1986). Model-based confidence intervals in mltipurpose surveys. *Applied Statistics*, 35, 276-280.

GRIMES, J.E., et SUKHATME, B.V. (1980). A regression-type estimator based on preliminary test of significance. *Journal of the American Statistical Association*, 75, 957-962.

HANSEN, M.H., et TEPPING, B.J. (1969). Progress and problems in survey methods and theory illustrated by the work of the United States Bureau of the Census. *New Developments in Survey Sampling*, (N.L. Johnson et H. Smith Jr., Éds.). New York: John Wiley & Sons.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

MARDIA, K.V., KENT, J.T., et BIBBY, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

MILLER, A.J. (1990). *Subset Selection in Regression*. London: Chapman and Hall.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications* (2ième éd.). New York: John Wiley & Sons.

ROYALL, R.M., et CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide* (Version 6, Vol. 2, 4ième éd.). Cary, NC: SAS Institute Inc.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SILVA, P.L.D.N. (1996). Some Asymptotic Results on the Mean Squared Error of the Regression Estimator Under Simple Random Sampling Without Replacement. Southampton: University of Southampton, Centre for Survey Data Analysis Technical Report 96-2.

Les résultats donnent à penser que, quand on estime la moyenne de la population d'une variable dépendante unique, les méthodes adaptatives proposées combinant l'estimateur de régression et une certaine forme de sélection d'un sous-ensemble de variables s'appuyant sur un estimateur approprié de l'erreur quadratique moyenne produisent un gain utile d'efficacité comparativement aux méthodes concurrentes. Cependant, ces stratégies risquent d'introduire un certain biais quand le pouvoir de prédiction des variables auxiliaires disponibles est faible et les estimateurs correspondants de l'erreur quadratique moyenne peuvent être biaisés considérablement, situation qui se traduit par une mauvaise couverture.

7. CONCLUSIONS ET ORIENTATIONS FUTURES

Nos résultats laissent entendre que, dans le cas de l'estimation par régression, on peut réaliser des gains d'efficacité en adoptant une méthode de sélection des variables fondées sur un des estimateurs de l'erreur quadratique moyenne v_d ou v_g. Dans le cas de la méthode de régression type, et compte tenu des renseignements limités fournis par la simulation, on ne dispose que de peu d'indices permettant de choisir entre ces deux estimateurs.

Les méthodes de sélection progressives d'un sous-ensemble de variables sont aussi efficaces que celles basées sur l'examen de tous les sous-ensembles possibles qui demandent beaucoup plus de calculs. Nos résultats indiquent également qu'il est possible d'améliorer la méthode de sélection d'un sous-ensemble par réduction du nombre de conditions quand on étudie une variable dépendante particulière.

Un des problèmes que pose la stratégie de sélection des variables tient au fait que l'estimation connexe de la variance est probablement biaisée pour l'estimation de l'erreur quadratique moyenne globale de l'estimateur de régression après la sélection des variables, donc, menant à une mauvaise couverture des intervalles de confiance types. Il convient donc de poursuivre les travaux en vue d'étudier d'autres méthodes d'estimation de la variance.

Le présent article porte principalement sur l'utilisation de l'estimation par régression en vue de réduire la variance d'échantillonnage dans le contexte de l'échantillonnage classique. En pratique, on utilise largement l'estimation par régression pour corriger les biais causés par les erreurs non dues à l'échantillonnage. Dans de telles applications, la question de savoir combien de variables auxiliaires il faut utiliser est également importante. Certaines variables pourraient être incluses pour des raisons indépendantes de l'erreur d'échantillonnage, par exemple, parce qu'on sait qu'elles sont des déterminants importants de la non-réponse. Néanmoins, la variance d'échantillonnage pourrait augmenter à mesure que croît le nombre de variables auxiliaires et, selon nous, il serait utile d'établir une règle de décision pour limiter le nombre de variables auxiliaires employées en tenant compte de la variance d'échantillonnage. En cas de biais non dû à l'échantillonnage, l'écart entre $\bar{x}$ et $\bar{X}$ sera généralement donné par $O_p(1)$ plutôt que par $O_p(n^{-1/2})$ et, par conséquent, les résultats présentés ici ne sont pas directement applicables. Il conviendra donc d'étudier l'extension de notre méthode à ce cas.

D'autres travaux doivent aussi être effectués pour étendre notre méthode aux plans de sondage complexes. Une méthode que l'on pourrait envisager pour les estimateurs généraux de régression, méthode considérée, par exemple, par Särndal et coll. (1992, sec. 6.4), consisterait à remplacer les poids g_i par les poids «généralisés», décrits par Särndal et coll. (1992, éq. 6.5.9), et à fonder la sélection des variables sur la minimisation de la version généralisée de v_g donnée par Särndal et coll. (1992, éq. 6.6.4).

Tableau 3
Matrice de corrélation pour les variables utilisées dans l'étude en simulation basée sur les données du Recensement de la population de 1988

Variable	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	0.23										
x_2	-0.04	0.2									
x_3	0.17	0.07	-0.40								
x_4	0.47	0.13	-0.15	0.12							
x_5	0.48	0.09	-0.11	0.15	0.83						
x_6	0.05	-0.09	-0.32	-0.03	0.22	0.20					
x_7	-0.17	0.01	-0.12	-0.01	-0.17	-0.31	0.16				
x_8	0.38	0.29	0.07	0.17	0.44	0.41	0.13	-0.20			
x_9	0.20	0.08	-0.06	0.04	0.30	0.25	0.16	-0.13	0.37		
x_{10}	0.43	0.23	0.33	0.17	0.39	0.39	-0.10	-0.30	0.49	0.26	
x_{11}	0.78	0.23	-0.00	0.22	0.54	0.54	0.01	-0.19	0.41	0.21	0.49

Les résultats de la simulation effectuée avec l'ensemble de dix variables auxiliaires ($x_1 - x_{10}$) sont présentés au tableau 2 ci-dessous. Comme prévu, ces résultats indiquent que les méthodes basées sur l'estimateur de régression sont plus efficaces que la méthode de la moyenne de l'échantillon. Cependant, les gains d'efficacité ne sont pas aussi importants que ceux indiqués au tableau 1, dans le cas de cinq variables auxiliaires ayant un plus grand pouvoir explicatif. Comme auparavant, les méthodes adaptatives basées sur la sélection progressive d'un sous-ensemble donnent des résultats comparables à ceux des méthodes fondées sur la sélection du meilleur sous-ensemble parmi tous les sous-ensembles possibles. De nouveau, les stratégies adaptatives utilisant v_d ou v_g comme estimateur de l'erreur quadratique moyenne sont un peu plus efficaces que les stratégies correspondantes basées sur v_s, quoique, dans ce cas, au prix d'une sous-couverture plus grande des intervalles de confiance nominaux de 95%.

Les stratégies adaptatives d'estimation les plus efficaces (Fd, Fg, Bd et Bg) mènent à une moyenne de la population et à une erreur quadratique moyenne entachées d'un biais non négligeable. En revanche, les stratégies FI et SS ne produisent aucun biais significatif pour la moyenne, même si l'estimation de l'erreur quadratique moyenne est entachée d'un certain biais dans le cas de la stratégie SS. Il convient de souligner le biais négatif important qui entachent les estimateurs de l'erreur quadratique moyenne, lequel est illustré par les différences entre les valeurs des colonnes intitulées EQM et MEEQM dans le tableau 2. Le pire biais semble être produit par les stratégies Fd, Fg, Bd et Bg. Viennent ensuite les stratégies Fs et Bs, puis les stratégies SS, FR et CN pour lesquelles le biais n'est pas si mauvais.

La comparaison de Fd et de Fg à CN montre que les deux premières méthodes aboutissent à un gain d'efficacité comparativement à la méthode de réduction du nombre de conditions, au prix d'une certaine augmentation du biais qui entache aussi bien l'estimateur de la moyenne que celui de l'erreur quadratique moyenne. Donc, même quand le pouvoir de prédiction des variables auxiliaires disponibles est grand, il est possible d'adopter une stratégie plus efficace que la stratégie CN.

Le choix d'un sous-ensemble fixe inapproprié (par exemple, le sous-ensemble saturé utilisé pour la stratégie SS) pourrait aboutir à des résultats médiocres au chapitre de l'efficacité et biaiser dans une certaine mesure l'estimation de l'erreur quadratique moyenne. Cependant, si, par exemple, on utilisait v_d au lieu de v_s comme estimateur de l'erreur quadratique moyenne dans le cas de la stratégie SS, on ne noterait aucun biais apparent (la MEEQM observée dans ce cas est 459.67, donc une valeur beaucoup plus proche de celle de l'estimation de l'erreur quadratique moyenne de la simulation, soit 462.71).

De nouveau, l'estimateur de régression ridge donne des résultats légèrement inférieurs à ceux de la stratégie du sous-ensemble saturé (SS), mais, cette fois-ci, sans aucun biais évident entachant l'estimation de la moyenne ou de l'erreur quadratique moyenne. Une fois de plus, la régression ridge s'avère plus efficace que la stratégie de réduction du nombre de conditions CN, mais l'écart entre les deux méthodes est plus faible ici. Elle donne également de bons résultats en ce qui concerne la couverture empirique de l'intervalle de confiance.

La stratégie FR donne de nouveau des résultats comparables à ceux des stratégies à sous-ensemble fixe FI et SS, donc, est surpassée par les stratégies fondée sur un estimateur de l'erreur quadratique moyenne de l'estimateur de régression tel que v_d ou v_g.

Tableau 2

Biais, erreur quadratique moyenne, moyenne des estimations de l'erreur quadratique moyenne et efficacité de diverses stratégies d'estimation de la moyenne de la variable dépendante y quand on dispose de dix variables auxiliaires ($x_1 - x_{10}$)

Stratégie d'estimation	BIAIS	EQM	MEEQM	Efficacité par rapport à SM (%)	Couverture[1] empirique (%)
SM) Moyenne de l'échantillon ($\bar{y}, v_s$)	0.25	620.09	619.05	100.00	91.8
Fs) Progressive ($\bar{y}_r, v_s$)	0.06	468.46	397.99	75.55	86.7
Fd) Progressive ($\bar{y}_r, v_d$)	-8.12	434.27	338.90	70.03	81.7
Fg) Progressive ($\bar{y}_r, v_g$)	-7.90	433.71	328.46	69.94	81.6
Bs) Meilleur ($\bar{y}_r, v_s$)	-0.00	466.16	397.59	75.18	86.6
Bd) Meilleur ($\bar{y}_r, v_d$)	-7.90	434.54	336.88	70.08	81.5
Bg) Meilleur ($\bar{y}_r, v_g$)	-7.60	433.26	326.05	69.87	81.6
FI) Fixe ($\bar{y}_r, v_s$)	0.45	490.49	461.86	79.10	89.0
SS) Saturé ($\bar{y}_r, v_s$)	-0.20	462.71	413.17	74.62	86.9
FR) PROC REG ($\bar{y}_r, v_s$)	-0.07	466.13	399.34	75.17	86.4
CN) Réd. Nbre. Cond. ($\bar{y}_r, v_s$)	3.49	562.91	450.36	90.78	87.3
RI) Ridge ($\bar{y}_{BC}, v_{DC}$)	1.05	480.18	472.82	77.44	89.4

[1] Couverture nominale de 95%.

Tableau 1

Biais, erreur quadratique moyenne, moyenne des estimations de l'erreur quadratique moyenne et efficacité de diverses stratégies d'estimation de la moyen*ne* de la variable dépendante y quand on dispose de cinq variables auxiliaires ($x_1 - x_4, x_{11}$)

Stratégie d'estimation	BIAIS	EQM	MEEQM	Efficacité rapport à SM (%)	Couverture[1] empirique (%)
SM) Moyenne de l'échantillon ($\bar{y}, v_s$)	0.25	620.09	619.05	100.00	91.8
Fs) Progressive ($\bar{y}_r, v_s$)	0.4	233.78	239.62	37.7	82.7
Fd) Progressive ($\bar{y}_r, v_d$)	-1.25	188.08	196.88	30.33	82.0
Fg) Progressive ($\bar{y}_r, v_g$)	-1.28	188.38	192.73	30.38	81.1
Bs) Meilleur ($\bar{y}_r, v_s$)	0.44	236.9	239.49	38.2	82.7
Bd) Meilleur ($\bar{y}_r, v_d$)	-1.22	190.52	196.84	30.72	82.0
Bg) Meilleur ($\bar{y}_r, v_g$)	-1.24	190.83	192.71	30.77	81.1
FI) Fixe ($\bar{y}_r, v_s$)	0.29	227.90	241.24	36.75	83.3
SS) Saturé ($\bar{y}_r, v_s$)	0.3	233.58	242.32	37.67	82.5
FR) PROC REG ($\bar{y}_r, v_s$)	0.38	235.86	240.26	38.04	82.5
CN) Réd. nbre. Cond. ($\bar{y}_r, v_s$)	0.34	507.33	483.63	81.82	89.8
RI) Ridge ($\bar{y}_{BC}, v_{DC}$)	2.12	304.95	257.07	49.18	82.5

[1] Couverture nominale de 95%.

résultats devraient, en principe, être meilleurs si on choisit un plus petit nombre de variables *x* pour les échantillons affichant les valeurs de c_g^2 les plus grandes, qui a laissé entrevoir cette propriété.

La comparaison avec la stratégie adaptative FR, basée sur la sélection type de sous-ensembles offerte par PROC REG de SAS, montre que le fait d'utiliser pour critère un estimateur approprié de l'erreur quadratique moyenne de l'estimateur de régression améliore les résultats. L'efficacité de FR est semblable à celle des méthodes traditionnelles basées sur un sous-ensemble fixe (FI-SS).

La faible efficacité de la méthode de sélection d'un sous-ensemble par réduction du nombre de conditions (CN) comparativement à celle de toutes les méthodes basées sur l'estimateur de régression est un résultat plus frappant, mais toutefois pas inattendu, car la méthode ne tient pas compte de la variable dépendante. Cette observation donne du poids à l'argument selon lequel, si la moyenne d'une variable dépendante particulière est la principale cible d'inférence, il convient d'en tenir compte quand on sélectionne les variables auxiliaires à utiliser avec l'estimateur de régression.

Dans le cas de l'ensemble de cinq variables auxiliaires, nous notons aussi que, pour chaque échantillon, la première variable éliminée pour réduire le nombre de conditions est l'approximation du revenu (x_{11}). Cette situation est due au fait que les valeurs propres (donc, le nombre de conditions) de la matrice CP dépendent des unités de mesure des variables auxiliaires. Puisque toutes les autres variables auxiliaires sont des dénombrements d'une sorte ou d'une autre, l'approximation du revenu est la variable affichant, de loin, la variance la plus forte. Son exclusion de chaque échantillon explique, en partie, la performance médiocre de cette méthode, puisqu'il s'agit du meilleur prédicteur unique de la variable dépendante.

Les travaux de Bankier ne font pas ressortir cette difficulté, car, dans le cas des données d'échantillon du Recensement de la population du Canada de 1991 auxquelles il applique sa méthode, toutes les variables auxiliaires étudiées sont des dénombrements de personnes, de familles ou de ménages, exprimées dans des unités comparables.

Contrairement aux valeurs propres de la matrice CP, l'estimateur de régression ne dépend ni de l'emplacement ni de la transformation d'échelle des variables auxiliaires. Donc, pour éviter que la méthode du nombre de conditions dépende arbitrairement des unités des variables auxiliaires, il est naturel de commencer par normaliser ces variables, puis de calculer le nombre de conditions de la matrice de corrélation de l'échantillon $\hat{R}_x$ plutôt que de la matrice $X_s^{*\prime} X_s^*$. Cependant, lors de l'essai de cette méthode, même le choix de valeurs modestes pour L (100) n'a abouti à l'élimination d'aucune variable auxiliaire. Par conséquent, l'ensemble saturé a été utilisé chaque fois, réduisant ainsi CN à SS.

La stratégie fondée sur l'estimateur de régression ridge (RI) donne de moins bons résultats que celle basée sur le sous-ensemble saturé (SS) en ce qui concerne l'efficacité. En outre, elle produit une estimation de l'erreur quadratique moyenne entachée d'un certain biais. Cette perte d'efficacité tient au fait que tous les poids doivent être supérieurs ou égaux à $1/N$, condition imposée uniquement dans ce cas-ci. En revanche, cette stratégie est beaucoup plus efficace que celle consistant à réduire le nombre de conditions (CN).

En ce qui concerne les taux de couverture empiriques, seule la stratégie de réduction du nombre de conditions CN donne des résultats proches de ceux obtenus par la méthode de la moyenne de l'échantillon (SM), chacune produisant un légère sous-couverture. Toutes les autres méthodes basées sur l'estimation par régression donne des taux de couverture comparables, nettement inférieurs à la cible de 95%.

FI) Sous-ensemble fixé de variables auxiliaires avec $(\bar{y}_r, v_s)$.

SS) Sous-ensemble saturé de variables auxiliaires avec $(\bar{y}_r, v_s)$.

FR) Sélection progressive d'un sous-ensemble au moyen de SAS PROC REG, avec $(\bar{y}_r, v_s)$.

CN) Sélection d'un sous-ensemble par réduction du nombre de conditions avec $(\bar{y}_r, v_s)$.

RI) Estimateur de régression ridge avec un sous-ensemble saturé de variables auxiliaires et un estimateur de variance, que nous représentons par v_{DC}, proposé par Dunstan et Chambers 1986), $(\bar{y}_{BC}, v_{DC})$.

Les stratégies Fs à Bg sont des variantes de deux méthodes que nous avons proposées pour sélectionner des sous-ensembles qui découlent de l'utilisation des trois estimateurs de l'erreur quadratique moyenne examinés à la section 3. Les stratégies FI et SS se fondent sur le même ensemble de variables auxiliaires quel que soit l'échantillon sélectionné. Dans le cas de SS, le sous-ensemble saturé englobant toutes les variables auxiliaires disponibles est utilisé constamment. Dans le cas de FI, nous avons choisi un sous-ensemble à partir de l'ensemble de cinq variables auxiliaires (x_1, x_4, x_{11} choisies) ou de dix variables auxiliaires ($x_1, x_2, x_5, x_8, x_{10}$ choisies) considéré, en appliquant une méthode de régression progressive type à l'ensemble de données sur la population. Puis, pour chaque échantillon, nous nous sommes servis du sous-ensemble sélectionné, d'où le nom de stratégie du «sous-ensemble fixé» pour FI. En pratique, cette stratégie n'est pas applicable, car on ne dispose pas des données sur la population pour la variable dépendante, mais nous la considérons théoriquement en tant que «meilleur scénario possible» dans le cadre de la méthode traditionnelle.

Dans le cas de la stratégie FR, nous utilisons «naïvement» SAS PROC REG pour exécuter la sélection progressive type d'un sous-ensemble pour chaque échantillon. La *valeur prédictive p* utilisée pour décider si une nouvelle variable devrait être incluse est la valeur implicite du programme, à savoir 0.50. Pour plus de détails, consulter SAS (1990, p. 1397).

Dans le cas de la sélection de sous-ensembles en vue de réduire le nombre de conditions CN, nous fixons à 1,000 la valeur du paramètre L qui contrôle la méthode. Pour la stratégie d'estimation par régression ridge RI, nous fixons à 1 la valeur de tous les coefficients de coût associés aux erreurs d'étalonnage qui entachent les diverses variables. Après avoir choisi la valeur de λ garantissant qu'aucun poids ne soit inférieur à $1/N$, nous rééchelonnons les poids de sorte que leur somme soient exactement 1, pour être certains que l'étalonnage soit exact au moment de l'estimation de la taille de la population.

Quelle que soit la stratégie utilisée, nous représentons par $\bar{y}(s)$ et $v[\bar{y}(s)]$ les estimations de la moyenne de population et de son erreur quadratique moyenne pour l'échantillon s, respectivement. Pour chaque stratégie, nous résumons les résultats de la simulation en calculant les estimations du biais, de l'erreur quadratique moyenne (EQM) et de la moyenne des estimations de l'erreur quadratique moyenne (MEEQM) pour l'ensemble des 1 000 échantillons répétés. Ces mesures sont données respectivement par

$$\text{BIAIS} = \sum_s [\bar{y}(s) - \bar{Y}]/1{,}000 \tag{12}$$

$$\text{EQM} = \sum_s [\bar{y}(s) - \bar{Y}]^2/1{,}000 \tag{13}$$

$$\text{MEEQM} = \sum_s v[\bar{y}(s)]/1{,}000. \tag{14}$$

Pour chaque stratégie, nous calculons également une mesure de l'efficacité en divisant l'erreur quadratique moyenne obtenue pour la simulation correspondante par l'erreur quadratique moyenne obtenue pour la simulation selon la stratégie de la moyenne de l'échantillon (stratégie SM) et en multipliant le résultat par 100. Nous calculons aussi, pour chaque stratégie les taux de couverture empiriques pour les intervalles de confiance de 95% fondés sur la théorie normale asymptotique. Ces taux, exprimés en pourcentage, sont présentés dans les dernières colonnes des tableaux 1 et 2.

Le Tableau 1 montre les résultats de la simulation visant à estimer la moyenne de la variable dépendante en se fondant sur l'ensemble formé des cinq variables auxiliaires ($x_1 - x_4, x_{11}$) ayant le pouvoir de prédiction le plus grand. Dans ce cas, l'utilisation de l'estimateur de régression améliore considérablement la précision de chaque stratégie d'estimation étudiée, sauf celle consistant à sélectionner un sous-ensemble par réduction du nombre de conditions (CN). Le biais est négligeable (inférieur à 1% en ce qui concerne le biais relatif absolu) pour toutes les stratégies d'estimation (la moyenne de y est 194.34), sauf, peut-être, la stratégie RI, pour laquelle on observe un léger biais.

Les résultats sont les mêmes pour les stratégies fondées sur la sélection progressive d'un sous-ensemble (Fs-Fg) et pour les stratégies correspondantes fondées sur la sélection à partir de tous les sous-ensembles possibles (Bs-Bg). Par conséquent, il est préférable d'appliquer les méthodes de sélection progressives, qui sont plus rapides et moins coûteuses.

Parmi les stratégies fondées sur la sélection progressive d'un sous-ensemble, Fd et Fg (avec v_d et v_g comme estimateur de l'erreur quadratique moyenne, respectivement) sont les plus efficaces et produisent des résultats fort semblables. Il convient aussi de souligner que Fd et Fg donnent de meilleurs résultats que FI et SS, à savoir les stratégies axées sur l'estimateur de régression s'appuyant sur un sous-ensemble fixe de cinq variables auxiliaires pour chaque échantillon. Cette observation est vraie dans le cas du sous-ensemble saturé (SS), ainsi que dans celui du sous-ensemble fixe choisi d'après des renseignements tirés de l'ensemble de la population (FI). Donc, il est possible d'obtenir de meilleurs résultats que ceux donnés par la méthode traditionnelle, c'est-à-dire l'utilisation d'un estimateur de régression avec un sous-ensemble fixé de variables auxiliaires, en adoptant une méthode adaptative consistant à choisir le «meilleur» estimateur de régression (sous-ensemble) pour chaque échantillon, du moins quand la variable dépendante étudiée est celle considérée pour la sélection du sous-ensemble. C'est la forte variation de la valeur c_g^2 d'un échantillon à l'autre, situation dans laquelle

Puis, nous avons sélectionné 1,000 échantillons de taille 100 à partir de cette population de simulation par échantillonnage aléatoire simple sans remise.

Avant d'examiner les résultats détaillés de la simulation, examinons quels avantages pourrait offrir la sélection des variables à la lumière de la discussion motivante fondée sur les modèles de la section 2. Rappelons que, comme le montre l'équation (4), si on se base sur le modèle (2), un terme c_g^2 gonfle la variance conditionnelle de $\bar{y}_r$ à cause de l'estimation de β. Nous avons évalué la distribution de c_g^2 sur les 1,000 échantillons, quand on considère cinq et dix variables auxiliaires. Dans le cas de cinq variables auxiliaires, la valeur de la médiane de c_g^2 est 0.036, celle du quartile supérieur est 0.056 et celle du maximum est 0.255. Ces résultats concordent grossièrement avec l'équation (5) qui implique que, en vertu du modèle, la valeur attendue de c_g^2 est $(1 - n/N)q/(n - q - 2) = 0.041$. Il convient de souligner que la forte variation de c_g^2 d'un échantillon à l'autre donne à penser qu'il pourrait être judicieux d'adopter une méthode consistant à sélectionner un ensemble distinct de variables pour chaque échantillon. La variation de c_g^2 est encore plus forte dans le cas de dix variables auxiliaires, la médiane étant de 0.078, le quartile supérieur de 0.107 et le maximum de 0.329, résultats qui concordent également grossièrement avec la valeur de 0.087 prévue par le modèle, conformément à l'équation (5). Cette interprétation dépend manifestement de la validité du modèle (2), laquelle est douteuse pour ces données, mais elle donne à penser que la sélection des variables permettrait de réaliser des gains d'efficacité.

Une autre façon d'évaluer les gains d'efficacité éventuels dus à la sélection des variables consiste à calculer des approximations de la variance de l'estimateur de régression en choisissant divers sous-ensembles des variables auxiliaires disponibles et en se servant de tous les enregistrements obtenus pour la population. La figure 1 représente graphiquement l'approximation que donne une version pour population finie de l'équation (5) calculée pour des sous-ensembles croissants des dix variables auxiliaires, où la variable ajoutée à chaque étape est celle qui produit la diminution la plus forte de l'approximation. Ces valeurs de l'approximation type de premier ordre axée sur le plan de sondage $(1 - f)S_e/n$ sont également représentées graphiquement à titre de référence, quoique, comme nous l'avons déjà souligné, cette approximation soit monotone, n'augmentant pas quand on ajoute de nouvelles variables auxiliaires. Les estimations par simulation de l'erreur quadratique moyenne de l'estimateur de régression correspondant à chaque sous-ensemble sont également représentées graphiquement. Le graphique montre clairement que si on choisit d'utiliser un estimateur de régression type contenant un ensemble fixe de variables auxiliaires, le sous-ensemble contenant cinq prédicteurs est le meilleur choix dans le cas de l'approximation normale de la variance fondée sur l'expression (5), tandis que le sous-ensemble saturé est plus indiqué dans le cas de l'approximation de la variance fondée sur le plan de sondage. Le graphique montre aussi que les estimations par simulation de l'erreur quadratique moyenne concordent davantage avec le modèle d'approximation normal qu'avec l'approximation type de premier ordre, particulièrement pour les grands sous-ensembles de variables auxiliaires. Nous obtenons des résultats comparables quand nous calculons les approximations correspondantes de la variance pour l'ensemble de cinq variables auxiliaires.

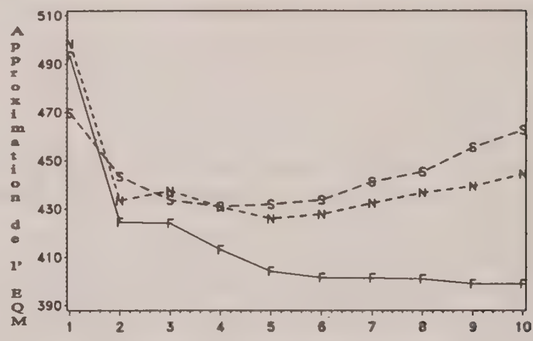


Figure 1. Approximations et estimations par simulation de l'EQM de l'estimateur de régression calculée, dans le cas d'une population finie, pour des sous-ensembles croissants de dix variables auxiliaires.

Donc, la distribution de c_g^2 obtenue par simulation de même que les approximations pour une population finie de la variance de l'estimateur de régression indiquent qu'on pourrait réaliser des gains d'efficacité en sélectionnant les variables pour cette population. Afin de déterminer si cette constatation s'applique à nos données, nous décrivons maintenant en détail l'étude en simulation.

Pour chaque échantillon répété (représenté par s) et pour chacun des deux sous-ensembles de variables auxiliaires considérés, nous avons calculé les estimations de la moyenne de population du revenu mensuel total, ainsi que les estimations correspondantes de la variance, en appliquant plusieurs stratégies d'estimation. Chaque stratégie est définie par la combinaison d'une méthode de sélection d'un sous-ensemble, d'un estimateur de la moyenne et d'un estimateur de la variance correspondante. Voici la liste des stratégies étudiées.

SM) Estimateur de la moyenne de l'échantillon, sans variables auxiliaires $(\bar{y}, v_s)$. Cette stratégie représente la norme à laquelle seront comparées toutes les autres.

Fs) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_s)$.

Fd) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_d)$.

Fg) Sélection progressive de variables auxiliaires avec $(\bar{y}_r, v_g)$.

Bs) Sélection du meilleur sous-ensemble possible de variables auxiliaires avec $(\bar{y}_r, v_s)$.

Bd) Sélection du meilleur des sous-ensemble possible de variables auxiliaires avec $(\bar{y}_r, v_d)$.

Bg) Sélection du meilleur sous-ensemble possible des variables auxiliaires avec $(\bar{y}_r, v_g)$.

à choisir un sous-ensemble γ^* dans Γ conformément à une règle déterminée par les données et par $\boldsymbol{S}$, l'estimateur ponctuel résultant étant $\bar{y}_r^{\gamma^*}$.

Pour chaque sous-ensemble déterminé γ, il s'ensuit, dans des conditions de régularité type (Isaki et Fuller 1982), que $\bar{y}_r^{\gamma}$ est convergent pour la moyenne de la population $\bar{Y}$, autrement dit que $\bar{y}_r^{\gamma} - \bar{Y} = o_p(1)$. Alors, pour une valeur donnée $\delta > 0$, $|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta$ implique que $|\bar{y}_r^{\gamma} - \bar{Y}| > \delta$ pour certains sous-ensembles γ. Nous pouvons donc écrire

$$\Pr(|\bar{y}_r^{\gamma^*} - \bar{Y}| > \delta) \leq \sum_{\gamma \in \Gamma} \Pr(|\bar{y}_r^{\gamma} - \bar{Y}| > \delta) \qquad (11)$$

et, comme Γ est fini, le membre droit de l'équation (11) converge vers zéro et il s'ensuit que $\bar{y}_r^{\gamma^*}$ est également convergent.

Cependant, la distribution de $\bar{y}_r^{\gamma^*}$ dépendra de la règle de sélection de façon complexe. Consulter Grimes et Sukhatme (1980) pour un examen de l'efficacité de $\bar{y}_r^{\gamma^*}$ dans le cas le plus simple où il existe seulement deux estimateurs possibles, à savoir un estimateur de régression à une seule variable x et un estimateur de différence (dont la moyenne est un cas particulier), et où les variables suivent conjointement une distribution normale.

Contrairement à $\bar{y}_r^{\gamma^*}$, il n'y a aucune raison que v^{γ^*} soit convergent pour $\mathrm{Var}(\bar{y}_r^{\gamma^*})$, même si v^{γ} est convergent pour $\mathrm{Var}(\bar{y}_r^{\gamma})$ pour chaque sous-ensemble γ déterminé. En particulier, nous pouvons nous attendre à ce que v^{γ^*} sous-estime $\mathrm{Var}(\bar{y}_r^{\gamma^*})$ si la règle de sélection est telle que v^{γ^*} représente le minimum de v^{γ}. Cet effet est semblable à la surestimation bien connue de R^2 après la sélection de sous-ensembles dans le cas de la régression linéaire multiple type (Miller 1990, p. 7 à 10).

6. ÉTUDE EN SIMULATION

Nous présentons ici une petite étude en simulation effectuée pour évaluer la performance des diverses méthodes de sélection des variables envisagées. Nous avons choisi comme population pour la simulation un ensemble de données comprenant 426 enregistrements obtenus pour les chefs de ménage qui ont répondu au questionnaire détaillé utilisé durant l'essai de recensement de population effectué en 1988 à Limeira, dans l'État de São Paulo, au Brésil.

Cet essai a été effectué à titre d'enquête pilote durant la préparation du Recensement de la population du Brésil de 1991. Le test comprenait deux vagues de collecte de données. Durant la première, chaque recenseur a visité tous les logements occupés dans une région de recensement donnée (une région comptant de 200 à 300 ménages, en moyenne) et a rempli un questionnaire abrégé. Ce dernier contenait quelques questions sur les caractéristiques du ménage et sur chaque membre du ménage (sexe, âge, lien avec le chef de ménage et niveau d'alphabétisme). Le questionnaire comprenait aussi, à l'intention des chefs de ménage uniquement, une question sur la scolarité et une autre sur le revenu mensuel total. Le revenu mensuel total déclaré par les chefs de ménage ne fournit qu'une approximation du revenu réel, en raison des limitations de la méthode d'interview suivie durant la première vague de collecte de données.

Puis, une deuxième vague de collecte de données a été entreprise dans chaque région de recensement. Les mêmes recenseurs ont rendu visite à un échantillon de un ménage sur dix (sélectionné systématiquement d'après la liste des logements occupés établie durant la première vague de collecte de données) pour recueillir des renseignements au moyen d'un questionnaire détaillé contenant toutes les questions du questionnaire abrégé et de nombreuses autres.

La taille de la population étudiée était de l'ordre de 44,000 ménages comptant en tout 188,000 personnes. La taille de l'échantillon correspondait à 10 % environ de la taille de la population. Pour limiter le coût du traitement informatique, nous avons utilisé, pour notre étude en simulation, une sous-population comprenant tous les enregistrements d'échantillon pour 426 chefs de ménage vivant dans 20 des 170 régions de recensement. Nous avons choisi ces enregistrements comme population pour la simulation parce qu'ils contiennent tous les renseignements détaillés fournis par le questionnaire utilisé pour interviewer les ménages faisant partie de l'échantillon, ainsi que les renseignements obtenus par personne interposée au moyen du questionnaire abrégé durant la première vague d'interviews.

Nous avons étudié le revenu mensuel total, tel qu'obtenu au moyen du questionnaire détaillé, à titre de variable dépendante principale (y), ainsi que 11 variables auxiliaires possibles, à savoir:

x_1 = indicateur du sexe du chef de ménage égal masculin;
x_2 = indicateur de l'âge du chef de ménage inférieur ou égal à 35;
x_3 = indicateur de l'âge du chef de ménage supérieur à 35 et inférieur ou égal à 55;
x_4 = nombre total de pièces dans le logement;
x_5 = nombre total de salles de bains dans le logement;
x_6 = indicateur de propriété du logement;
x_7 = indicateur de ce que le type de logement est une maison;
x_8 = indicateur de la possession d'au moins une voiture par le ménage;
x_9 = indicateur de la possession d'un téléviseur couleur par le ménage;
x_{10} = nombre d'années d'études du chef de ménage;
x_{11} = approximation du revenu mensuel total du chef de ménage.

À partir de ces 11 variables, nous avons construit deux ensembles distincts de variables auxiliaires pour nos simulations. Nous avons défini le premier ensemble en prenant cinq variables auxiliaires, nommément $x_1, ..., x_4$ et x_{11} possédant un pouvoir de prédiction de y raisonnable, particulièrement à cause de la présence de l'approximation du revenu x_{11}. Le deuxième ensemble que nous avons examiné contenait dix variables auxiliaires, nommément $x_1, ..., x_{10}$, ayant, à cause de l'exclusion de x_{11}, un pouvoir de prédiction plus faible que l'ensemble précédent. À titre de référence, nous présentons au tableau 3 en annexe la matrice de corrélation de la population pour la variable observée y et pour les 11 variables auxiliaires de la population.

choisir au départ la moyenne de l'échantillon comme estimateur, puis à ajouter la variable qui minimise l'estimation de l'erreur quadratique moyenne. La procédure est répété jusqu'à ce que l'estimation de l'erreur quadratique moyenne commence à augmenter, le sous-ensemble de variables qui produit l'estimation minimale de l'erreur quadratique moyenne étant sélectionné à ce moment-là.

Nous pouvons comparer ces deux méthodes à une autre, inspirée des travaux de Bankier et de ses associés (voir Bankier (1990) et Bankier, Rathwell et Majkowski (1992) que nous appelons *méthode de réduction du nombre de conditions*. Avant de la décrire, notons que l'on peut aussi exprimer l'estimateur de régression donné en (1) par

$$\bar{y}_r = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \qquad (9)$$

où X_s^* est la matrice $n \times (q+1)$ pour laquelle $x_i^{*\prime} = (1, x_{i1}, ..., x_{iq})' = (1 \vdots x_i')$ représente la i-ième rangée, $\bar{x}^* = (1 \vdots \bar{x}')'$ et $\bar{X}^* = (1 \vdots \bar{X}')'$ représentent, respectivement, les vecteurs des moyennes de l'échantillon et de la population des x_i^* et y_s est le vecteur $n \times 1$ des observations de la variable dépendante dans l'échantillon.

Donc, l'estimateur de régression dépend de l'inversion de la matrice $X_s^{*\prime}X_s^*$ de produits croisés, matrice dont les conditions deviennent parfois inappropriées et qui, par conséquent, gonfle la variance de l'estimateur de régression.

Bankier (1990) propose une méthode en deux étapes pour calculer les estimateurs de régression des moyennes (ou des totaux) en vertu de laquelle on élimine certaines colonnes de la matrice de données auxiliaires X_s^* afin de réduire le nombre de conditions de la matrice de produits croisés $X_s^{*\prime}X_s^*$ et d'éviter ainsi les situations indésirables (poids négatifs ou aberrants, caractéristiques rares ou dépendance linéaire exacte entre colonnes). Bankier et ses collaborateurs (1992) décrivent en détail l'application de la méthode aux données du Recensement de la population du Canada de 1991. Il convient de souligner que, si elle intègre la sélection des variables, la méthode élaborée par Bankier et ses collaborateurs ne vise pas à être efficace pour une variable observée particulière. Elle est axée principalement sur l'étalonnage, avec, en parallèle, la fourniture d'un ensemble unique de poids applicable à toutes les variables observées.

Nous pouvons décrire notre méthode de réduction du nombre de conditions au moyen de l'algorithme ci-après, qui se fonde sur l'élimination à rebours des variables auxiliaires qui produisent un grand nombre de conditions pour la matrice de produits croisés $CP = X_s^{*\prime}X_s^*$, au lieu de l'ajout progressif de variables décrit par Bankier et coll. (1992).

1) Calculer la matrice de produits croisés $CP = X_s^{*\prime}X_s^*$ en tenant compte de toutes les colonnes existantes au départ (sous-ensemble saturé).
2) Calculer la forme canonique d'Hermite de CP, soit H (voir Rao 1973, p. 18), et déceler les singularités en examinant les éléments diagonaux de H. Tout élément diagonal nul de H indique que les colonnes correspondantes de $X_s^{*\prime}X_s^*$ (et X_s^*) dépendent linéairement d'autres colonnes (voir Rao 1973, p. 27). Éliminer chacune de ces colonnes en supprimant les rangées et colonnes correspondantes de $X_s^{*\prime}X_s^*$.
3) Après avoir éliminé toute colonne linéairement dépendante, calculer le nombre de conditions $c = \lambda_{max}/\lambda_{min}$ de la matrice CP réduite, où λ_{max} et λ_{min} représentent la plus grande et la plus petite valeurs propres de CP, respectivement. Si $c < L$ est une valeur précise, arrêter et utiliser toutes les variables auxiliaires restantes.
4) Sinon, procéder à l'élimination à rebours comme suit. Pour chaque k, supprimer les k-ième rangée et colonne de CP, et calculer de nouveau les valeurs propres et le nombre de conditions de la matrice réduite. Calculer les réductions du nombre de conditions $r_k = c - c_k$, où c_k est le nombre de conditions après avoir éliminé les k-ième rangée et colonne de CP. Déterminer $r_{max} = \max_k(r_k)$ et $k_{max} = \{k : r_{max} = r_k\}$, et éliminer la colonne k_{max} en supprimant les rangée et colonne k_{max} de CP. Poser $c = c_{k_{max}}$ et itérer tant que $c \geq L$ et que $q \geq 2$, en commençant chaque nouvelle itération par la matrice CP réduite résultant de la précédente.

Une autre méthode que nous étudions est l'estimation par régression ridge de Bardsley et Chambers (1984). Au lieu de se fonder sur la sélection de sous-ensembles des variables auxiliaires existantes, elle s'appuie sur le relâchement des propriétés d'étalonnage de l'estimateur de régression en vue d'obtenir des estimations plus stables. L'estimateur de régression ridge est représenté par

$$\bar{y}_{BC} = [n\bar{y} + (N\bar{X}^* - n\bar{x}^*)'(\lambda C^{-1} + X_s^{*\prime}X_s^*)^{-1}X_s^{*\prime}y_s]/N \qquad (10)$$

où λ est un paramètre scalaire de «ridging» et où C est une matrice diagonale de coefficients de «coût» associés aux erreurs d'étalonnage tolérées quand on estime les totaux des variables auxiliaires au moyen de $\bar{y}_{BC}$.

Bardsley et Chambers (1984) laissent entendre qu'on pourrait utiliser la spécification de la matrice C pour contrôler l'influence de chaque variable auxiliaire sur l'estimateur résultant de la moyenne de la variable dépendante, donc imiter le procédé de sélection des sous-ensembles. Comme dans le cas du paramètre de «ridging» λ, ils proposent de choisir la valeur la plus petite, de façon qu'aucun poids de cas implicites ne soit inférieur à $1/N$ (ou à 1 pour les totaux estimés).

5. PROPRIÉTÉS DES ESTIMATEURS DE RÉGRESSION APRÈS LA SÉLECTION DES VARIABLES

Dans le cas des méthodes fondamentales de sélection des variables, nous considérons un ensemble de stratégies d'estimation $\boldsymbol{S} = \{(\bar{y}_r^{\gamma}, v^{\gamma}); \gamma \in \Gamma\}$, où $\bar{y}_r^{\gamma}$ et v^{γ} représentent l'estimateur de régression et un estimateur de sa variance, respectivement, pour un sous-ensemble γ des q variables auxiliaires existantes, et où Γ représente l'ensemble des sous-ensembles. La procédure de sélection des variables consiste

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}(1 - n/N + c_g^2) \tag{4}$$

où c_g est le coefficient de variation des g_i de l'échantillon.

Pour étudier la corrélation prévue entre c_g^2 et q, nous étendons maintenant le modèle en supposant que les variables x_i sont indépendantes et distribuées identiquement selon la loi normale. Notant que $(\bar{x} - \bar{X})$ et $\hat{S}_x$, sont des quantités indépendantes et que $E_M(\bar{y}_r - \bar{Y} \mid x_i) = 0$, nous obtenons la variance non conditionnelle

$$\begin{aligned} &\mathrm{Var}_M(\bar{y}_r - \bar{Y}) \\ &= \sigma^2 n^{-1}\{1 - n/N + \mathrm{tr}[E_M[(\bar{X} - \bar{x})(\bar{X} - \bar{x})']E_M(\hat{S}_x^{-1})]\} \\ &= \sigma^2 n^{-1}(1 - n/N)[1 + q/(n - q - 2)] \end{aligned} \tag{5}$$

en nous appuyant sur le fait que $n^{-1}\hat{S}_x^{-1}$ est caractérisé par une distribution de Wishart inverse (Mardia, Kent et Bibby 1979, p. 69 et 85). Ce résultat est également vérifié pour les grandes valeurs de n, même si les conditions normales ne sont pas respectées, en ce sens que $[1 - n/N + c_g^2]/(1 - n/N)[1 + q/(n - q - 2)]$ continue de converger vers 1 quand n augmente et que q est constant (dans des conditions faibles).

L'expression (5) rend la corrélation avec q explicite. À mesure que q augmente, nous pouvons nous attendre à ce que σ^2 diminue, mais que $E_M(c_g^2)$ augmente. La diminution de σ^2 deviendra faible après l'inclusion de quelques variables x importantes, donc la variance commencera à augmenter au moment où le nombre de variables x représentera une fraction non négligeable de la taille de l'échantillon.

Les résultats (4) et (5) se fondant sur de fortes hypothèses de modélisation ne nous fournissent que la motivation. Dans le cas général, $\bar{x} - \bar{X} = O_p(n^{-1/2})$ (distribution de randomisation avec conditions types de régularité), de sorte que le dernier terme de (3) est d'ordre $O_p(n^{-2})$. Quand le modèle (2) ne doit pas être vérifié, on obtient une approximation asymptotique de deuxième ordre plus générale de l'erreur quadratique moyenne de $\bar{y}_r$ correspondant au plan en généralisant le théorème 4.1 de Deng et Wu (1987). Le lecteur trouvera des précisions dans Silva (1996).

Notre objectif consistant à élaborer une méthode de sélection des variables qui réduise au minimum l'erreur quadratique moyenne estimée de $\bar{y}_r$ nous allons maintenant examiner les estimateurs de cette erreur quadratique moyenne.

3. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE DE L'ESTIMATEUR DE RÉGRESSION MULTIPLE

Nous obtenons un estimateur simple de l'erreur quadratique moyenne de $\bar{y}_r$ en généralisant l'expression (7.29) de Cochran (1977, p. 195) au cas de plusieurs variables auxiliaires, soit:

$$v_s = \frac{1 - f}{n}\hat{S}_e \tag{6}$$

où $\hat{S}_e = (n - q - 1)^{-1}\sum_{i \in s}\hat{e}_i^2$ et où $\hat{e}_i = (y_i - \bar{y}) - (x_i - \bar{x})'b$.

Toutefois, cet estimateur ne tient pas compte de l'élément $O(n^{-2})$ de l'erreur quadratique moyenne. Donc, en guise de deuxième estimateur de l'erreur quadratique moyenne, nous généralisons l'estimateur v_d étudié par Deng et Wu (1987) au cas où q est général. Il s'agit d'un cas particulier de l'estimateur G_2 de la variance fondé sur un modèle et résistant au biais qui a été proposé au départ par Royall et Cumberland (1978) pour traiter le cas où les variances résiduelles produites par le modèle (2) sont constantes. Cet estimateur est donné par

$$v_d = \frac{1 - f}{n(n - 1)}\sum_{i \in s}\alpha_i\hat{e}_i^2 \tag{7}$$

où

$$\alpha_i = (g_i^2 - 2g_i f + f)/\{(1 - f)[1 - (x_i - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})/(n - 1)]\}.$$

Nous supposions au départ que v_d serait un estimateur de deuxième ordre non biaisé, comme Deng et Wu (1987, éq. 4.4) le démontre dans le cas où $q = 1$. Malheureusement, cette hypothèse n'est pas confirmée dans le cas général où $q > 1$, mais on peut s'attendre à ce que le biais entachant v_d soit plus petit que celui de v_s, comme l'indiquent les expressions du biais de deuxième ordre obtenues pour v_s et v_d par Silva (1996).

Le problème que pose l'utilisation de v_d comme estimateur de la variance tient à ce qu'il est difficile de généraliser son application aux plans de sondage complexes. Donc, nous envisageons comme troisième estimateur de la variance une version modifiée d'un estimateur proposé par Särndal, Swensson et Wretman (1989), définie par:

$$v_g = \frac{1 - f}{n(n - q - 1)}\sum_{i \in s}g_i^2\hat{e}_i^2. \tag{8}$$

En principe, cet estimateur devrait se comporter de façon similaire à v_d puisque $\alpha_i = g_i^2 + O_p(n^{-1/2})$. Conformément à la terminologie utilisée par Särndal et coll. (1992, p. 232), les g_i sont les poids g appropriés dans les conditions d'échantillonnage aléatoire simple si on adopte (2) comme modèle sous-jacent. L'expression (8) diffère de l'estimateur correspondant proposé par Särndal et coll. (1989, exemple 4.4) en ce sens que nous utilisons le dénominateur $(n - q - 1)$ plutôt que le dénominateur original $(n - 1)$.

4. MÉTHODES DE SÉLECTION DES VARIABLES

Nous considérons deux méthodes fondamentales de sélection des variables. Premièrement, nous examinons une méthode englobant *tous les sous-ensembles* qui consiste à calculer un des estimateurs de l'erreur quadratique moyenne v_s, v_d ou v_g de la section 3 pour les 2^q sous-ensembles possibles des q variables auxiliaires (incluant toujours la coordonnée à l'origine) et à choisir le sous-ensemble qui correspond à l'estimation la plus faible de l'erreur quadratique moyenne. Cette méthode comporte manifestement de longs calculs si q est grand. Donc, nous considérons une deuxième méthode, appelée méthode de *sélection progressive*, qui consiste à

expliquent «la plus grande partie» du R^2 de l'échantillon (consulter Särndal et coll. 1992, sec. 7.9.1). Pourtant, un plus grand soutien théorique semble nécessaire, particulièrement quand le nombre de variables x est grand.

Une raison supplémentaire d'examiner plus formellement le problème de sélection des variables tient à ce qu'une telle étude apporterait des éclaircissements quant à l'incidence éventuelle de la sélection des variables sur l'inférence. On reconnaît de longue date le fait que la sélection d'estimateurs d'après l'échantillon peut avoir des effets sur les propriétés des estimateurs choisis (Hansen et Tepping 1969, App.), mais rares sont les études qui visent à déterminer quels pourraient être ces effets.

Dans le présent article, nous examinons une méthode de sélection des variables visant à minimiser l'erreur quadratique moyenne de $\bar{y}_r$. Toutefois, à la section 2, nous commençons par étudier la corrélation entre l'erreur quadratique moyenne de $\bar{y}_r$ et le nombre de variables x, puis, à la section 3, nous considérons d'autres estimateurs de l'erreur quadratique moyenne de $\bar{y}_r$. Enfin, à la section 4, nous présentons certaines méthodes de sélection des variables fondées sur ces estimateurs.

Nous comparons notre méthode de sélection des variables à quatre méthodes existantes. En premier lieu, nous considérons la méthode habituelle consistant à utiliser un sous-ensemble fixe de variables auxiliaires, indépendamment de l'échantillon observé. Puis, nous examinons une «méthode de réduction du nombre de conditions» inspirée des travaux de Bankier (1990), selon laquelle on élimine des variables auxiliaires afin de réduire le nombre de conditions d'une matrice donnée de produits croisés des variables x.

En troisième lieu, nous suivons Bardsley et Chambers (1984) et considérons une méthode de régression ridge. Plutôt que d'inclure une sélection de variables, cette méthode vise à résoudre le problème éventuel de multicollinéarité de l'estimateur de régression grâce à une modification de l'estimateur, en tenant compte de certaines erreurs d'étalonnage. La méthode de régression ridge et la méthode de réduction du nombre de conditions présentent toute deux l'avantage de ne pas obliger à préciser la variable dépendante y, car elles visent à produire un ensemble unique de poids «d'étalonnage» applicable à toutes les variables étudiées. Cependant, leur application ne garantit pas que l'efficacité augmentera. Dans les tableaux présentés à la section 6, les résultats obtenus par ces méthodes sont séparés des autres par une ligne pour indiquer qu'ils sont différents.

En quatrième lieu, nous envisageons la sélection des variables en s'appuyant sur les critères des tests de signification classiques. Selon nous, dans l'ensemble, l'objectif de la sélection des variables lors d'estimation des paramètres de populations finies par régression est assez différent de celui de l'estimation des paramètres ou de la prédiction des valeurs de y dans le cas des observations isolées produites par la régression classique (Miller 1990). Néanmoins, il nous paraît souhaitable de prendre cette dernière comme point de comparaison.

À la section 5, nous examinons les propriétés de l'estimateur de régression après sélection des variables d'après les variances estimées. À la section 6, nous décrivons une étude empirique effectuée en vue de comparer les méthodes de sélection des variables que nous proposons aux méthodes concurrentes décrites plus haut. Cette étude porte sur les données d'un recensement pilote effectué dans la municipalité de Limeira, au Brésil, en prévision du Recensement de la population du Brésil de 1991. À la section 7, nous présentons nos conclusions et certaines orientations que pourraient prendre les travaux de recherche futurs.

2. RELATION ENTRE LA VARIANCE DE L'ESTIMATEUR DE RÉGRESSION ET LE NOMBRE DE VARIABLES x

Commençons par définir certaines conventions de notation. Représentons par $U = \{1,...,N\}$ une population finie de N éléments distincts et par $s \subset U$ un échantillon de n éléments distincts tirés de U selon un plan d'échantillonnage aléatoire simple sans remise. Supposons que $x_i = (x_{i1},...,x_{iq})'$ est le vecteur $q \times 1$ des variables auxiliaires associées au i-ième élément de la population. Nous supposons qu'on connaît les valeurs des $x_i\,(i \in s)$ de l'échantillon, ainsi que le vecteur des moyennes de la population $\bar{X} = N^{-1}\sum_{i\in U} x_i$. Le vecteur des moyennes de l'échantillon est représenté par $\bar{x} = n^{-1}\sum_{i\in s} x_i$.

Représentons par y_i la valeur de la variable étudiée y pour le i-ième élément de la population et supposons qu'on n'observe les valeurs de y_i que pour $i \in s$. Nous visons à estimer la moyenne de la population $\bar{Y} = N^{-1}\sum_{i\in U} y_i$.

L'estimateur de régression de $\bar{Y}$ est donné par $\bar{y}_r$ dans l'équation (1), où $\bar{y} = n^{-1}\sum_{i\in s} y_i$, $b = \hat{S}_x^{-1}\hat{S}_{xy}$, $\hat{S}_x = n^{-1}\sum_{i\in s}(x_i - \bar{x})(x_i - \bar{x})'$, et $\hat{S}_{xy} = n^{-1}\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})$.

Cet estimateur peut être amené par le modèle linéaire sous-jacent

$$y_i = \beta_0 + x_i'\beta + \varepsilon_i \tag{2}$$

où les ε_i sont des perturbations indépendantes dont les moyennes sont nulles et dont la variance commune est σ^2, puisque nous pouvons écrire $\bar{y}_r = \hat{\beta}_0 + \bar{X}'\hat{\beta}$, où $\hat{\beta}_0 = \bar{y} - \bar{x}'b$ et $\hat{\beta} = b$ représentent les estimateurs par les moindres carrés de β_0 et β, respectivement. En vertu de ce modèle, la variance de $\bar{y}_r - \bar{Y}$ conditionnelle à x_i peut s'écrire

$$\mathrm{Var}_M(\bar{y}_r - \bar{Y} \mid x_i) = \sigma^2 n^{-1}[1 - n/N + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})]. \tag{3}$$

On peut interpréter le dernier terme comme l'effet de l'estimation de β au moyen de b. À mesure que le nombre q de variables x augmente, on peut s'attendre à ce que la variance résiduelle σ^2 diminue, alors que le terme $(\bar{X} - \bar{x})'\hat{S}_x^{-1}(\bar{X} - \bar{x})$ pourrait augmenter à mesure que $\hat{S}_x^{-1}$ devient plus instable. Une autre façon d'interpréter ce terme consiste à représenter $\bar{y}_r$ comme un estimateur pondéré $\bar{y}_r = n^{-1}\sum_{i\in s} g_i y_i$ où $g_i = 1 + (\bar{X} - \bar{x})'\hat{S}_x^{-1}(x_i - \bar{x})$ Alors, nous pouvons écrire (3) comme suit

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 25-35
Statistique Canada

Sélection des variables pour l'estimation par régression dans le cas des populations finies

PEDRO L. D. NASCIMENTO SILVA et CHRIS J. SKINNER[1]

RÉSUMÉ

Les auteurs examinent la sélection des variables auxiliaires pour l'estimation par régression des paramètres des populations finies dans le cas d'un plan de sondage aléatoire simple. Ce problème fondamental que posent les méthodes d'échantillonnage fondé sur un modèle ou assisté par un modèle prend une importance d'ordre pratique quand le nombre de variables disponibles est grand. Les auteurs élaborent une méthode consistant à minimiser un estimateur de l'erreur quadratique moyenne, puis, la comparent à d'autres en utilisant un ensemble fixe de variables auxiliaires, un test de signification classique, une méthode de réduction du nombre de conditions et une méthode de régression ridge. Selon les résultats de l'étude, la méthode proposée est efficace. Les auteurs soulignent que la méthode de sélection des variables influe sur les propriétés des estimateurs types de la variance, ce qui entraîne par conséquent un problème d'estimation de la variance.

MOTS CLÉS: Informations auxiliaires; étalonnage; enquêtes par sondage; sélection d'un sous-ensemble; régression ridge.

1. INTRODUCTION

L'estimation par régression est une méthode utilisée fréquemment dans le cas des enquêtes par sondage pour intégrer des informations auxiliaires sur la population $\bar{X}$ (Cochran 1977, chap. 7). Dans le cas élémentaire où on connaît la moyenne de la population d'un vecteur de variables x_i et qu'on effectue un échantillonnage aléatoire simple, l'estimateur de régression de la moyenne de la population $\bar{Y}$ d'une variable étudiée y_i prend la forme

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})'b \tag{1}$$

où $\bar{y}$ et $\bar{x}$ sont respectivement les moyennes d'échantillon de y_i et de x_i, et où b est le vecteur d'échantillon des coefficients de régression linéaire de y_i sur x_i.

L'estimation par régression est utile pour au moins trois raisons. Premièrement, la méthode est souple. En principe, on peut intégrer dans $\bar{X}$ n'importe quel nombre de moyennes de population de variables continues ou binaires. Plus précisément, la stratification a posteriori se dégage comme un cas particulier (Särndal, Swensson et Wretman 1992, sec. 7.6). La méthode s'étend aussi au traitement de plans de sondage complexes. Deuxièmement, l'estimation par régression présente certaines propriétés d'optimisation de l'efficacité. À cet égard, consulter, par exemple, Isaki et Fuller (1982, Theorem 3). Troisièmement, $\bar{y}_r$ présente la propriété d'«étalonnage» voulant que, si y_i est une des variables de x_i de sorte que $\bar{Y}$ est connue, alors $\bar{y}_r = \bar{Y}$ (Deville et Särndal 1992).

Dans le présent document, nous cherchons à déterminer comment sélectionner les variables x utilisées dans l'estimateur de régression. Cette question présente un intérêt pour deux raisons. La première, d'ordre pratique, tient simplement au fait que, dans certaines circonstances, le nombre de variables x_i peut être très grand. Par exemple, dans le cas du recensement de la population de plusieurs pays, on enregistre les valeurs de certaines variables au moyen d'un «questionnaire abrégé» rempli par tous les membres de la population et les valeurs d'autres variables, au moyen d'un «questionnaire détaillé» rempli par un échantillon de la population. Donc, on connaît les moyennes de population des variables visées par le questionnaire abrégé, ainsi que leurs carrés, leurs cubes, leurs produits et ainsi de suite. Habituellement, on connaît aussi la définition des petites régions. Donc, la dimension de x_i en tant que vecteur contenant des fonctions des variables visées par le questionnaire abrégé ainsi que des variables fictives représentant chaque petite région pourrait facilement être de l'ordre de plusieurs milliers. Le cas échéant, la sélection de variables x devient une nécessité pratique.

La deuxième raison est surtout fondamentale dans le cas d'une méthode d'échantillonnage assisté par modèle ou fondé sur un modèle. Dans le contexte de l'estimation par régression, on peut décrire ces méthodes comme suit. Pour commencer, on choisit un modèle de régression possédant un «bon pouvoir de prédiction», de sorte que l'estimateur de régression ait une «bonne efficacité». Puis, on adopte une méthode d'inférence fondée sur le plan de sondage dans le cas de la méthode assistée par modèle (Särndal et coll. 1992) ou une méthode de prédiction basée sur un modèle, dans le cas de la méthode fondée sur un modèle. Si de nombreux articles ont été publiés sur le problème d'inférence, les chercheurs semblent avoir accordé officiellement très peu d'attention au problème de sélection du modèle. En pratique, il semble que les efforts se limitent à choisir les variables x «principales» qui

[1] Pedro L. D. Nascimento Silva, IBGE-Departamento de Metodologia, Avenida Chile 500, Rio de Janeiro-RJ, Brasil; et Professor Chris J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, UK.

fonctionner, notons $k_0 <<< N$. C'est le cas que nous illustrons dans le tableau B. Dans le tableau B, nous avons limité notre attention à une seule valeur de N, N =5,000 grappes, bien que les résultats puissent être élargis facilement.

Tableau B
Pr{l'échantillon inverse choisit le motif (1,1, ..., 1)}

k_0	k_0/N	$M = 10$	$M = 100$
2	.0004	0.9998	0.9998
5	.001	0.9982	0.998
10	.002	0.9919	0.9911
20	.004	0.9663	0.9627
30	.006	0.9245	0.9166
40	.008	0.8687	0.8553
50	.01	0.8015	0.7821

Évidemment, à mesure que k/N devient plus petit, un échantillon systématique représente un meilleur inverse approximatif. Seule l'expérience pourrait confirmer si l'approximation à $k_0 = 20$ et $k_0/N = 0{,}004$, par exemple, est adéquate. Nous estimons qu'elle pourrait l'être, surtout du fait que l'utilisation d'un inverse systématique entraîne normalement des calculs de la variance plus prudents (car typiquement la corrélation intra-grappe [$\rho > 0$] est supérieure à 0).

BIBLIOGRAPHIE

BELLHOUSE, D. (1988). A brief history of random sampling methods. *Handbook of Statistics,* 6, 1-14.

CLEVELAND, W. (1993). *Visualizing Data.* Summit, NJ: Hobart Press.

COCHRAN, W. (1977). *Sampling Techniques.* New York: Wiley.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7, 139-172.*

FELLEGI, I. (1980). Approximate tests of independence et goodness of fit based on multistage samples. *Journal of the American Statistical Association,* 75, 261-268.

HANSEN, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 162-179.

HINKINS, S., OH, H.L., et SCHEUREN, F. (1995). Using an Inverse Algorithm for Testing of Independence Based on Stratified Samples. George Washington University, Rapport Technique.

HUGHES, S., MULROW, J., HINKINS, S., COLLINS, R., et UBERALL, B. (1994). Section 3, *Statistics of Income – 1991, Corporation Income Tax Returns*, 9-17. Washington, DC: Internal Revenue Service.

KISH, L. (1995). The Hundred Years Wars of Survey Sampling. Centennial representative Sampling Conference, Rome, le 31 mai 1995.

LAHIRI, D. (1951). A method for sample selection providing unbiased ratio estimates, *Bulletin de l'Institut International de Statistique*, 34, 72-86.

McCARTHY, P., et SNOWDEN, C. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*. Series 2, No. 95, DHHS Pub. No. (PHS) 85-1369. Washington, DC: Public Health Service.

MULROW, J., et SCHEUREN, F. (1996). Measuring to improve quality and productivity in a processing environment. *Data Quality*, 2, 11-20.

OSBORNE, D., et GAEBLER, T. (1992). *Reinventing Government.* New York: Plume.

PFEFFERMANN, D., et NATHAN, G. (1985). Problems in model identification based on data from complex samples. *Bulletin de l'Institut International de Statistique*, 68.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

SÄRNDAL, C.-E., et SWENSSON, B. (1993). Présentation à la Washington Statistical Society sur la nature changeante du paradigme de l'échantillonnage.

SCHEUREN, F. (1972). Topics in Multivariate Finite Population Sampling and Data Analysis. George Washington University Dissertation de doctorat.

SCHEUREN, F. (Éd.) (1995). *What is a Survey?* tiré d'une série de brochures publiée par l'American Statistical Association pour accroître la connaissance des enquêtes.

SKINNER, C., HOLT, D., et SMITH, T., (Éds.) (1989). *Analysis of Complex Surveys.* New York: Wiley.

WESTFALL, P., et YOUNG, S. (1993). *Resampling-Based Multiple Testing.* New York: Wiley.

WOLTER, K. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

inverses. Cette façon de procéder semble extrêmement gauche.

Nous avons indiqué que, dans certains cas, il n'est peut-être pas trop difficile de rééchantillonner plusieurs fois à l'aide de l'algorithme inverse afin d'atteindre une efficacité raisonnable. Mais que dire du cas où l'usager d'un échantillon stratifié s'intéresse à des sous-populations? Si les domaines d'intérêt sont en réalité les strates, il n'est pas avantageux pour l'usager d'utiliser les échantillons aléatoires simples produits à l'aide de l'algorithme inverse. Si les domaines d'intérêt chevauchent les strates et s'ils sont petits, le nombre d'échantillons requis pour l'algorithme inverse risque d'être très grand si l'on veut maintenir une estimation raisonnable pour les domaines.

Enfin, mentionnons brièvement un autre problème que nous avons examiné. De nombreux plans de sondage à plusieurs degrés retiennent réellement une seule unité primaire d'échantillonnage par strate. Les strates sont alors appariées à des fins d'estimation de la variance. Nous avons déjà noté qu'il existe un inverse pour cette approximation que l'on peut rendre à peu près aussi valable que l'est cette approximation à l'origine. Y a-t-il moyen d'obtenir une meilleure approximation à l'aide de la stratégie inverse directement?

Derniers mots – Notre profession évolue appréciablement. La révolution mondiale de la qualité a certainement eu des répercussions (Mulrow et Scheuren 1996). Nous assistons à une refonte de la façon de mener des enquêtes, de la conception à la collecte des données et à l'utilisation de ces données par la clientèle. Le présent exposé représente peut-être un modeste apport à cet égard.

REMERCIEMENTS

Nous tenons à remercier les arbitres et le rédacteur adjoint de leurs remarques pertinentes et de leur dévouement. Le document original que nous avons présenté n'était qu'une ébauche de l'article publié. Nous tenons également à remercier Phil Kott qui a décrit nos travaux en cours lors de différentes réunions de la Washington Statistical Society.

ANNEXE

Supposons un échantillon de k grappes d'une population de N grappes, dans laquelle chaque grappe compte le même nombre d'unités, M. Dans l'algorithme de sondage inverse, la première étape consiste à choisir le vecteur $(m_1, m_2, ..., m_k)$ qui contient le nombre d'unités à choisir de chaque grappe. Notons q le nombre de valeurs non zéro de m_i. La probabilité de sélection du motif avec $q = k$, c'est-à-dire le motif avec $m_i = 1$, pour tous les $i = 1, 2, ..., k$, est la suivante:

$$\Pr(q = k) = M^{k-1} \frac{(N-1)(N-2)...(N-k+1)}{(NM-1)(NM-2)...(NM-k+1)}.$$

Appelons cette probabilité P_1. Si $NM >>> k$ il est possible d'établir une approximation de P_1 à l'aide de

$$\prod_{i=1}^{k-1} \frac{(N-i)}{N} = \frac{(N-1)(N-2)...(N-k+1)}{N^{k-1}}.$$

Considérons ensuite la partition de k correspondant à $q = k - 1$; cela correspond exactement à une partition de k, c'est-à-dire $\{1,1, ..., 1,2\}$. Il existe $k(k-1)$ motifs également probables de $(m_1, ..., m_k)$ avec $q = k - 1$. La probabilité de sélection d'un vecteur $\boldsymbol{m}$ avec $q = k - 1$ est la suivante:

$$\Pr(q = k - 1) = \frac{k(k-1)(M-1)}{2M(N-k+1)} P_1.$$

Par conséquent, il n'est pas difficile de calculer la probabilité que le $\boldsymbol{m}$ choisi compte soit $q = k$ ou $q = k - 1$. Le tableau ci-dessous donne quelques exemples des deux valeurs de M.

Tableau A
$\Pr(q = k - 1 \text{ ou } q = k)$

k	N	$M = 10$	$M = 100$
4	8	.92	.90
4	20	.99	.98
10	20	.38	.34
10	30	.63	.59
10	50	.83	.80
10	200	.99	.98
50	500	.35	.30
50	1000	.70	.66
50	5000	.98	.98

Pour k petit, il n'est pas difficile de calculer l'entière distribution des probabilités nécessaires à la production de $\boldsymbol{m}$. Mais à mesure que k augmente, le nombre de partitions augmente et ce calcul devient difficile ou du moins fastidieux. Pour $k = 4$, il n'existe que 4 partitions; pour $k = 10$ il existe 39 partitions possibles. On peut voir dans le tableau A que, à mesure que l'échantillon en grappes devient «plus grand», si le taux d'échantillonnage est assez faible, c'est-à-dire si $k << N$, on pourrait se limiter à calculer les probabilités pour ces deux partitions de façon à inverser approximativement l'échantillon en grappes. Pour $k = 10$ et $N = 200$, ces deux partitions tiennent compte essentiellement de toute la distribution des probabilités.

La probabilité de sélection d'une seule unité par grappe ($q = k$) est plus faible que les valeurs du tableau A; donc, pour utiliser un inverse systématique, nous voudrions $k <<< N$. On peut l'obtenir dans certains contextes lorsque le nombre de grappes est grand et que nous sommes prêts à prendre k très petit, quitte à rééchantillonner l'enquête originale de façon répétée comme il a été décrit à la section 3.

Pour illustrer, supposons un échantillon de taille k_0 où, bien entendu, $k_0 < k$, de façon qu'un inverse soit possible. De plus, afin de vérifier si un inverse systématique pourrait

À force de rééchantillonner de 500 à 1,000 fois, on a réduit la variance au même ordre de grandeur que l'échantillon stratifié. Même avec 100 sous-échantillons, nous avons de bons résultats ici, ce qui indique que le recours à un algorithme inverse pourrait donner de bons résultats pour ce genre de strate. Il n'est pas recommandé pour autant qu'un algorithme inverse soit utilisé en général avec un si faible rééchantillonnage. Sans doute, dans des populations très asymétriques, il en faudrait un bien plus grand nombre.

4. APPLICATIONS POSSIBLES ET PROCHAINES ÉTAPES

Dans le présent exposé, nous avons montré qu'il existe des algorithmes de plan de sondage inverses dans certains cas spéciaux. Nous n'avons toujours pas de résultat général, à supposer qu'il en existe un. Il s'agit là clairement d'un élément du problème qu'il y a lieu d'approfondir. Comme la plupart des outils, un algorithme de sondage inverse n'est pas nécessairement le meilleur choix dans certains cas; ce n'est peut-être même pas un choix raisonnable dans certaines circonstances. Il existe toutefois des applications dans lesquelles il semble offrir des avantages et il y a donc lieu de le considérer. Dans la présente section, nous décrivons brièvement des domaines dans lesquels cette méthode pourrait être utile et nous abordons certaines limites et certains problèmes qui subsistent.

Perspective axée sur la clientèle – Il est bon de souligner que notre stratégie est axée sur la clientèle. Même s'il n'était pas possible de les justifier pour d'autres raisons, on pourrait envisager les algorithmes inverses dans le cadre de la «réinvention» (p. ex. Osborne et Gaebler 1992). À l'heure actuelle, de nombreuses enquêtes complexes d'envergure ne sont peut-être pas suffisamment utiles pour la société, pour des raisons de sous-analyse grave ou même d'analyse fautive:

– À long terme, nous devons chercher à rendre la clientèle actuelle et éventuelle plus sensible aux enquêtes et à la quantification en général, par exemple à l'aide de la nouvelle série *What Is a Survey*? (sous la direction de Scheuren 1995).
– À court terme, nous devons commencer là où se trouve notre clientèle, en tenant compte du rôle souvent minime que jouent les données d'enquête dans leur processus de prise de décisions. Il y a certainement lieu de songer à réduire le fardeau cognitif que notre clientèle doit porter pour utiliser nos «produits» d'enquête complexes.

Choix de possibilités – Les gens sont de plus en plus sensibilisés aux faiblesses du paradigme de randomisation traditionnel (p. ex. Särndal et Swensson 1993). Mentionnons en particulier tout le travail que nous devons accomplir pour corriger les erreurs non imputables à l'échantillonnage. Cet aspect est exprimé dans Rao et Shao (1993). En reprenant les rajustements possibles pour ces erreurs non imputables à l'échantillonnage dans le cadre d'un échantillonnage aléatoire simple, nous pourrions peut-être même progresser davantage.

Depuis des décennies, les praticiens des enquêtes élaborent des plans de sondage extrêmement complexes pour ensuite en tirer des estimations d'intervalle de confiance et des estimations ponctuelles efficaces. Que savons-nous réellement des distributions que nos estimateurs d'échantillon produisent lorsque la taille utile des échantillons est petite à modérée? Serons-nous en mesure de bénéficier pleinement de la «révolution de visualisation» qui se produit actuellement (p. ex. Cleveland 1993)? Et en particulier en présence d'erreurs non imputables à l'échantillonnage? Nous devrions peut-être procéder de façon à toujours examiner les distributions. Le recours à un algorithme d'échantillonnage inverse serait une possibilité parmi d'autres (voir aussi Pfeffermann et Nathan 1985). De toute façon, des outils de visualisation plus puissants pour les enquêtes complexes pourraient aider même les vétérans parmi nous à approfondir leurs intuitions et à mieux les rattacher à la population particulière à l'étude. Évidemment, les efforts de visualisation entraînent également une baisse des prix d'utilisation des données d'enquête.

Un problème épineux qui se prêterait peut-être à un algorithme de sondage inverse est le cas où nous avons un plan de sondage à deux unités primaires d'échantillonnage par strate, avec L strates, où L est petit, par exemple moins de 30. Supposons également que, pour quelques-unes des variables de l'enquête, la stratification et la répartition en grappes n'ont pas d'importance, c'est-à-dire que l'effet du plan est $\delta = 1$, approximativement. Pour ces variables, ne serait-il pas possible de rendre la stabilité de l'estimation de la variance plus grande à l'aide de la méthode de rééchantillonnage qu'à l'aide de la méthode à répétition compensée que l'on applique normalement à l'estimation de la variance?

Un autre exemple que nous examinons est le cas où l'utilisateur s'intéresse à des tests d'indépendance en tableaux 2×2, fondés sur des données d'échantillon stratifié (Hinkins, Oh et Scheuren 1995). Pour la statistique du test du chi carré, nous en sommes à la comparaison de nos résultats à la stratégie proposée par Scheuren (1972) et Fellegi (1980). Il semble jusqu'à présent que la puissance de notre méthode soit comparable à celle de ces stratégies plus familières (comme on pourrait s'y attendre, par exemple, de Westfall et Young, 1993). C'est peut-être un cas où le travail supplémentaire exigé par l'algorithme d'échantillonnage inverse offrirait de réels avantages, en plus de simplement rendre le recours à des outils familiers plus facile pour les usagers, ceux-ci pouvant examiner la distribution et non pas seulement une valeur p.

Exemples de problèmes qui subsistent – Nous présentons ici des exemples de problèmes qui subsistent relativement à notre algorithme inverse. Ainsi, que se passe-t-il lorsque nous ignorons la taille de la population? Que se passe-t-il lorsque la population comporte plus d'une unité élémentaire, les personnes par exemple pour une analyse, les ménages pour une autre, les quartiers pour une troisième? Il existe des réponses pour ces difficultés, mais elles nous semblent ponctuelles. Dans de nombreuses enquêtes, par exemple, nous devinons la valeur N et nous utilisons cette valeur dans la stratification a posteriori. Ce degré d'approximation serait peut-être acceptable pour un inverse. Quant au problème des unités d'analyse multiples, nous pourrions mener plusieurs

$$t_{**} = \frac{1}{g}\sum_{j=1}^{g} t_j = \frac{1}{g}\sum_{j=1}^{g} N\bar{x}_j = \frac{1}{g}\sum_{j=1}^{g}\frac{N}{m}\sum_{i=1}^{m} x_{ji}$$

$$s_j^2 = \left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(x_{ji} - \bar{x}_j)^2$$

$$s_*^2 = \left(\frac{1}{gm-1}\right)\sum_{j=1}^{g}\sum_{i=1}^{m}(x_{ji} - \bar{x}_{**})^2$$

$$\text{où} \quad \bar{x}_{**} = \frac{t_{**}}{N} = \frac{1}{gm}\sum_j\sum_i x_{ji}.$$

À noter que la variance de l'échantillon fondée sur toutes les gm unités peut s'exprimer comme suit:

$$s_*^2 = \frac{1}{mg-1}\left[(m-1)\sum_{j=1}^{g} s_j^2 + \frac{m}{N^2}\sum_{j=1}^{g}(t_j - T)^2 - \frac{mg}{N^2}(t_{**} - T)^2\right].$$

Par conséquent

$$E(s_*^2) = \frac{1}{mg-1}\left[g(m-1)S^2 + \frac{m}{N^2}\sum_{j=1}^{g}\mathrm{Var}(t_j) - \frac{mg}{N^2}\mathrm{Var}(t_{**})\right].$$

La réécriture donne

$$\mathrm{Var}(t_{**}) = N^2\left(\frac{m-1}{m}\right)S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{mg-1}{mg}\right)E(s_*^2).$$

Donc, en remplaçant S^2 et $\mathrm{Var}(t_i)$ par des estimations non biaisées et en remplaçant $E(s_*^2)$ par s_*^2, nous pouvons produire des estimations à peu près non biaisées de $\mathrm{Var}(t_{**})$.

Il est peut-être bon de souligner que ce résultat n'exige pas que l'usager connaisse quoi que ce soit au sujet du plan de sondage original. Si l'on indique aux usagers comment inverser le plan de sondage original, ils pourront, à l'aide de sous-échantillonnages répétés, presque atteindre l'efficacité du plan de sondage original et déterminer aisément les erreurs d'échantillonnage appropriées. Une condition s'applique à ce résultat, à savoir que la taille du sous-échantillon doit être telle que $m \geq 2$. Soit dit en passant, pour $m = 2$, l'expression de la variance devient:

$$\mathrm{Var}(t_{**}) = \frac{N^2}{2}S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{2g-1}{2g}\right)E(s_*^2).$$

Par conséquent, comme ci-dessus, il serait possible de construire un estimateur de la variance pour des plans de sondage à deux unités primaires d'échantillonnage par strate.

3.3 Illustration SDR

Nous examinons ici un exemple d'algorithme inverse et son fonctionnement. Notre point de départ est l'échantillon des sociétés SDR (statistique des revenus). Comme nous l'avons noté antérieurement, l'échantillon SDR comporte essentiellement un plan de sondage de type échantillon aléatoire simple stratifié et il est donc possible de l'inverser (sous-section 2.2).

Nous sommes d'avis que de nombreux usagers SDR trouveront un échantillon aléatoire simple inverse complet plus utile et plus facile à utiliser que la base de données d'échantillons stratifiés au complet. Un objectif provisoire pourrait être de leur fournir un ensemble d'échantillons aléatoires simples. Un système plus souple serait de fournir le logiciel interactif permettant à l'usager de désigner les échantillons aléatoires simples qui l'intéressent, choisis à même la base de données tout entière.

Dans nos simulations, nous avons utilisé quatre des strates dans l'échantillon SDR des déclarations des sociétés, c'est-à-dire les strates représentant les plus petites sociétés ordinaires (Hughes, Mulrow et coll., 1994). Comme l'indique le tableau 1, l'échantillon stratifié (de quatre strates) comportait 15,618 unités et le plus grand échantillon aléatoire simple qui peut être choisi est $m = 2,224$. Le tableau montre également la taille des populations et la variance estimative de la variable Actif total, à l'intérieur de chaque strate.

Tableau 1
Taille de la population de sociétés et de l'échantillon, ainsi que la variance estimative des strates, pour quatre strates SDR

Strate (h)	N_h	n_h	S_h^2 (en milliers)
1	1,376,801	3,889	222,808
2	552,909	2,224	670,162
3	678,371	4,005	12,796,578
4	436,023	5,500	14,984,753

La variable Actif total a été utilisée parce qu'il s'agit de la principale variable de stratification; par conséquent, la perte de précision qu'entraînerait l'ignorance de la stratification serait relativement grande. En effet, c'est ce que l'on a observé.

On trouvera ci-dessous le rapport de la variance du total estimatif en fonction de g échantillons aléatoires simples, de 2,224 chacun, divisé par la variance du total fondée sur l'échantillon stratifié. Le tableau présente des valeurs de g entre 1 et 1,000. Par exemple, si un seul échantillon aléatoire simple est choisi, la variance du total estimatif est 29 fois plus grande que la variance du total stratifié.

g	Augmentation relative de la variance
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

l'échantillonnage à l'intérieur de chaque strate, nous pourrions employer un ou plusieurs des inverses exacts ou approximatifs de façon à obtenir deux sélections EAS à l'intérieur de chaque strate. Afin d'obtenir un échantillon aléatoire simple global, nous utiliserions l'algorithme inverse de la sous-section 2.3 pour ces deux sélections et, en fin de compte, nous n'aurions que deux sélections globalement.

2.5.4 Quelques remarques au sujet des plans de sondage à plusieurs degrés

Dans la présente sous-section, nous avons rapidement abordé quelques plans de sondage à plusieurs degrés et nous avons fourni des inverses exacts ou approximatifs. Les résultats ont été obtenus grâce à des résultats antérieurs des sous-sections 2.3 et surtout 2.4. Bien entendu, de nombreux plans de sondage à plusieurs degrés ne correspondent à aucun des cas spéciaux examinés, notamment ceux qui comportent des sélections systématiques au dernier degré.

De nombreux lecteurs se demanderont peut-être comment une méthode qui choisit uniquement un échantillon de taille deux (comme nous l'avons fait à la sous-section 2.5.3) peut être de quelque utilité pratique. La prochaine section apportera peut-être des éclaircissements.

3. RÉÉCHANTILLONNAGE EN VUE D'UNE PUISSANCE ACCRUE

3.1 Contexte général

Le prélèvement d'un échantillon aléatoire simple unique et plus petit à partir d'un échantillon plus complexe et plus grand peut convenir dans certaines situations. Toutefois, pour la plupart des usagers, la perte de puissance entre l'estimation qui se fonde sur l'échantillon complexe et l'estimation qui se fonde sur un échantillon aléatoire simple est inacceptable.

Afin d'augmenter la puissance de notre stratégie, nous avons tout naturellement considéré des techniques de rééchantillonnage. La taille des échantillons aléatoires simples qu'il est possible de prélever est limitée, mais nous pouvons répéter le processus. En répétant la procédure de sous-échantillonnage dans son ensemble, nous pouvons produire g échantillons aléatoires simples ayant chacun la taille m, où chaque échantillon aléatoire simple est choisi indépendamment à même l'échantillon original global. Chaque répétition doit inclure toutes les étapes de la procédure de sous-échantillonnage. Ainsi, dans le cas stratifié, la taille des sous-échantillons de strate doit être redéfinie à l'aide de la distribution hypergéométrique.

Dans la présente section, nous indiquons des conditions dans lesquelles la précision des estimations fondées sur des échantillons aléatoires simples multiples peut être rapprochée arbitrairement de la précision des estimations originales. Pour commencer, définissons notre notation.

Soit D, tout plan de sondage inversible (p. ex. un plan du type abordé à la section 2). Soit T, la quantité de la population qui nous intéresse (p. ex. un total de population); soit T_D, un estimateur non biaisé de T calculé à partir de l'échantillon $\boldsymbol{S}_D$. Supposons g échantillons aléatoires simples qui sont prélevés indépendamment de l'échantillon $\boldsymbol{S}_D$ donné et notons t_i l'estimateur de l'i-ième échantillon aléatoire simple. On peut dès lors montrer que

$$\text{si } E(t_i \mid S_D) = T_D$$

$$\text{alors } \operatorname{Var}\left(\frac{1}{g}\sum_{i=1}^{g} t_i\right) = \operatorname{Var}(T_D) + \frac{1}{g}\left(\operatorname{Var}(t_1) - \operatorname{Var}(T_D)\right).$$

Preuve: Puisque les g répétitions du processus d'échantillonnage aléatoire simple sont conditionnellement indépendantes, nous avons

$$\text{pour } i \neq j,\ E(t_i t_i \mid \boldsymbol{S}_D) = T_D^{\,2}.$$

Par conséquent, inconditionnellement, pour i non égal à j,

$$\begin{aligned}\operatorname{Cov}(t_i, t_j) &= E(t_i t_j) - T^2 \\ &= \operatorname{Var}(T_D).\end{aligned}$$

Et le résultat suit directement.

Quelques-unes des conditions de cette preuve peuvent être assouplies; si T_D est biaisé, on peut obtenir des résultats semblables pour l'EQM au lieu de la variance. Toutefois, la condition

$$E(t_i \mid S_D) = T_D$$

est nécessaire. Or cette condition n'est pas satisfaite pour les estimateurs de rapport. Par contre, si la condition est satisfaite séparément pour le numérateur et pour le dénominateur de l'estimateur de rapport, et si la taille finale de l'échantillon combiné est suffisamment grande pour qu'une approximation de la série de Taylor soit acceptable, il est possible de trouver des résultats semblables pour des approximations de la variance pour des rapports de la façon normale. Soit dit en passant, même dans le plan de sondage de deux unités primaires d'échantillonnage par strate, cette stratégie est valable, pourvu que nous puissions obtenir une estimation non biaisée de chaque échantillon individuel de taille 2. Pour des estimations de totaux, ce peut être le cas, à supposer que l'on puisse construire un inverse à chaque degré d'échantillonnage.

3.2 Estimation de l'erreur d'échantillonnage pour des moyennes ou des totaux

Par voie de rééchantillonnage, il est possible d'obtenir presque la même précision que l'estimateur du plan original. Mais puisque les échantillons aléatoires simples rééchantillonnés ne sont que conditionnellement indépendants, l'estimation de l'erreur-type n'est pas aussi simple que si un seul échantillon aléatoire simple avait été prélevé. Toutefois, l'estimation demeure relativement simple.

Soit S^2, la variance de la population pour la variable X; soit T, son total de population. Pour les moyennes, les totaux et les variances d'échantillons calculés à partir des échantillons aléatoires simples produits, notons

De nombreux plans en grappes ne se rapportent à aucun des cas particuliers examinés. Pour quelques-uns d'entre eux, nous supposons qu'il n'existe peut-être pas d'algorithmes inverses exacts. En particulier, le cas général d'un échantillonnage de type probabilité proportionnelle à la taille sans remise semble être un tel cas, y compris la variante souvent utilisée d'une probabilité proportionnelle à la taille sans remise systématique. Cela peut être ou ne pas être un problème pour les praticiens qui utilisent souvent l'hypothèse (normalement) prudente qu'il s'agit d'un échantillonnage avec remise; dans ce cas il existerait un algorithme inverse selon le même ordre d'approximation supposé dans l'estimation des variances.

2.5 Plans de sondage en grappes à plusieurs degrés

Qu'en est-il des plans de sondage à plusieurs degrés? Peut-on les inverser? Dans certains cas, nous croyons que la réponse est oui. Trois plans de sondage seront considérés: (1) un plan à deux degrés avec échantillonnage aléatoire simple aux degrés 1 et 2 (sous-section 2.5.1); (2) un plan faisant appel à un échantillonnage de type probabilité proportionnelle à la taille pour le premier degré et un échantillonnage aléatoire simple pour le deuxième (sous-section 2.5.2); (3) le plan de sondage très important à degrés multiples stratifiés avec deux unités primaires d'échantillonnage par strate, qui mérite au moins un bref aperçu.

Comme nous le verrons, il est possible d'élargir assez facilement les résultats stratifiés et à un degré. Pour le confirmer, notre stratégie de base consiste à appliquer de façon répétée les méthodes déjà décrites.

2.5.1 Plans de sondage à plusieurs degrés avec échantillonnage aléatoire simple pour les deux degrés

Supposons, tout d'abord, qu'à l'origine un échantillon aléatoire simple (EAS) de k grappes, toutes de taille M, a été prélevé au premier degré et qu'un sous-échantillon aléatoire simple de taille «r» a été prélevé au deuxième degré, dans chaque grappe retenue au premier degré.

Comme antérieurement, notre échantillon inverse ne peut pas être plus grand que k. Supposons d'abord que $1/(NM - k + 1)$ est à peu près égal à $1/NM$ et nous pouvons dès lors utiliser un algorithme inverse de type échantillon aléatoire simple avec remise, puisque EASar et EASsr sont très proches l'une de l'autre. À l'aide des résultats de la sous-section 2.4.3, nous prélevons un EASar de k grappes et ensuite, dans chaque grappe retenue, nous prenons une observation au hasard. Une autre possibilité serait, comme à la sous-section 2.4.1, de déterminer d'abord le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$. Une fois les m_i déterminés, un échantillon aléatoire simple sans remise de taille m_i est choisi dans la grappe $i, i = 1, 2, ..., k$. Il peut s'agir d'un résultat presque exact, sauf qu'il est possible que l'échantillon inverse de deuxième degré de taille m_i soit plus grand que l'échantillon original de deuxième degré de taille «r». Si cela se produit, nous pouvons tout de même faire appel aux résultats de la sous-section 2.4.2 et prélever notre échantillon de deuxième degré avec des «paramètres substituables». Dans ce deuxième cas, l'échantillon réel obtenu ne serait plus fixe, mais il resterait conditionnellement un échantillon aléatoire simple. Si les grappes de premier degré sont de taille inégale mais prélevées avec remise, nous pouvons encore une fois avoir recours à la stratégie utilisée à la sous-section 2.4.2 et créer des «paramètres substituables». La taille des échantillons est aléatoire et nous n'obtenons un inverse de type EAS que conditionnellement.

Une autre façon de procéder est de noter que l'échantillon aléatoire simple le plus grand qui peut être choisi à l'aide d'un algorithme inverse est de taille $k_0 = \min\{k, r\}$. Il s'agit d'abord de déterminer le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$, où la somme des m_i doit maintenant être k_0 au lieu de k. Une fois les m_i déterminés, un échantillon aléatoire simple de taille m_i est choisi dans la grappe $i, i = 1, 2, ..., k$. La distribution de probabilités à utiliser pour choisir les m_i est la suivante:

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1} \cdots \binom{M}{i_k}}{\binom{NM}{k_0}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}$$

où $0 \le i_j \le k_0$, $i_1 + i_2 + ... + i_k = k_0$, et q est le nombre de i_j non zéro.

Notons enfin, pour les grappes de taille tant égale qu'inégale, qu'il semble exister la possibilité d'un inverse systématique approximatif, moyennant bien sûr, plus ou moins, les mêmes réserves décrites ci-dessus.

2.5.2 Plans de sondage à plusieurs degrés avec échantillonnage PPT au premier degré et EAS au deuxième degré

Encore une fois, notre échantillon inverse ne peut pas être plus grand que k. Il est évident qu'une façon de construire un inverse serait d'utiliser les résultats de la sous-section 2.4.3. Plus particulièrement, nous prélèverions un échantillon aléatoire simple avec remise de k grappes et, ensuite, une observation au hasard dans chaque grappe retenue. Il est possible que d'autres algorithmes inverses existent également. Un inverse systématique semble raisonnable, pourvu que la probabilité de sélection de la même grappe plus d'une fois soit faible à presque nulle.

2.5.3 Plans de sondage à plusieurs degrés stratifiés avec deux unités primaires d'échantillonnage par strate

Est-il possible d'inverser des plans à deux UPÉ (unité primaire d'échantillonnage)? Notre réponse est oui, si les sélections intra-strate se font de l'une ou l'autre façon discutée en détail antérieurement. C'est essentiellement le seul cas que nous abordons.

Compte tenu de nos résultats des sous-sections 2.3 et 2.4, il est évident que pour qu'il existe un inverse, la taille de l'échantillon m ne saurait être supérieure à $m = 2$. Suivant

$$\frac{\binom{NM_* - M_+}{k - k_0}}{\binom{NM_*}{k}}$$

et tous les échantillons de taille k_0 ont la même chance d'être choisis à l'aide de l'algorithme inverse.

Il existe malheureusement une probabilité positive que cette stratégie entraîne la sélection d'un échantillon sans éléments. Cela pourrait se produire s'il y avait une différence importante dans la taille des grappes. Toutefois, lorsque le nombre de grappes k dans l'échantillon original est grand, il est peu probable que cela soit un problème.

Comme dans le cas des tailles de grappes égales, on a accès à une approximation en utilisant un sous-échantillon systématique comme inverse. Cette fois-ci, nous souhaitons avoir un pas d'échantillonnage au moins aussi grand que la taille de grappe maximale. Soit dit en passant, le recours à un inverse systématique offrirait l'avantage d'un meilleur contrôle de la taille réelle du sous-échantillon prélevé.

2.4.3 Échantillonnage en grappes à un degré, à grappes inégales, à probabilités inégales

Si l'on choisit un échantillon de k grappes avec PPT, il est possible qu'il existe un algorithme inverse. Supposons que les échantillons sont choisis avec remise d'une population constituée de N grappes, la taille des grappes $M_1, M_2, ..., M_N$. étant inégale. Supposons également que la mesure de la taille est soit égale à M_i ou proportionnelle à M_i. Dès lors, à chaque prélèvement, on a:

$$\text{Pr(sélection grappe } j) = \frac{M_j}{M_+} \qquad \text{où } M_+ = \sum_{i=1}^{N} M_i. \tag{13}$$

Enfin, puisque l'on prélève un échantillon à un degré, une fois la grappe j choisie toutes les unités M_j de cette grappe sont comprises dans l'échantillon.

Un algorithme inverse dans ce cas-ci devrait entraîner un EAS avec remise. Autrement dit, pour tout vecteur $\boldsymbol{S}$ résultant de k sélections indépendantes de la population, la probabilité de sélection du vecteur ordonné est la suivante:

$$\text{Pr(sélection } \boldsymbol{S}) = \left(\frac{1}{M_+}\right)^k. \tag{14}$$

Un algorithme inverse consiste à simplement choisir au hasard une unité de chaque grappe dans l'échantillon en grappes. Puisque les grappes ont été choisies avec remise, il convient de considérer les grappes prélevées comme étant ordonnées, dans l'ordre de leur sélection, ou dans un ordre fixe quelconque. Si donc la population contient 20 grappes, un échantillon en grappes possible de taille $k = 5$ est (7, 5, 7, 18, 6), *etc.*

La population est constituée de M_+ unités, notées $u_1, u_2, ..., u_{M+}$. Soit S, un échantillon donné avec remise, $S = (s_1, s_2, ..., s_k)$, soit $\boldsymbol{c} = (c_1, c_2, ..., c_k)$ la grappe associée pour chaque unité. Par exemple, supposons la population suivante:

Unités	Grappe
1	$u_1\ u_2\ u_3\ u_4$
2	$u_5\ u_6\ u_7\ u_8$
3	$u_9\ u_{10}\ u_{11}$
4	$u_{12}\ u_{13}\ u_{14}$
5	$u_{15}\ u_{16}\ u_{17}$
6	$u_{18}\ u_{19}\ u_{20}$

et $k = 3$. Dès lors l'échantillon $(s_1 = u_2, s_2 = u_4, s_3 = u_{17})$ correspond à $\boldsymbol{c} = (1, 1, 5)$. L'échantillon $(s_1 = u_{18}, s_2 = u_{19}, s_3 = u_{18})$ correspond à $\boldsymbol{c} = (6, 6, 6)$. À noter que ce deuxième échantillon ne peut être choisi que si la grappe 6 est la seule grappe choisie dans l'échantillon en grappes.

Pour un échantillon S donné de taille k, et pour le vecteur $\boldsymbol{c}$ correspondant de membres, de la grappe, la probabilité inconditionnelle de sélection de S à l'aide de l'algorithme inverse est la suivante:

$$\text{Pr(sélec. } S \mid \text{échan. en gr. } \boldsymbol{c}) * \text{Pr(sélec. } \boldsymbol{c}) = \left(\prod_{i=1}^{k} \frac{1}{M_{c(i)}}\right)\left(\prod_{i=1}^{k} \frac{M_{c(i)}}{M_+}\right) \tag{15}$$

qui est égale à la probabilité souhaitée, équation (14).

À noter que ce même algorithme inverse est valable lorsque k grappes sont choisies selon une probabilité proportionnelle à la taille avec remise, mais qu'un échantillon de taille fixe m est choisi (échantillon aléatoire simple sans remise) de la grappe retenue, en supposant que $M_i > m$ pour toutes les grappes i.

2.4.4 Quelques remarques au sujet des plans de sondage à un degré

Nous avons vu que, avec un peu de prudence, il est possible de construire des algorithmes inverses pour plusieurs cas particuliers dans lesquels l'échantillon original comporte un plan en grappes à un degré. Deux de nos résultats s'appliquent à des échantillons en grappes tirés à probabilités égales sans remise. Le troisième est un plan de type probabilité proportionnelle à la taille avec remise.

Il est même possible qu'un inverse systématique pratique soit valable à titre d'approximation de l'algorithme inverse correct lorsque nous avons un échantillon en grappes. L'approximation est valable lorsqu'on utilise un échantillon aléatoire simple avec remise qui «se rapproche» d'un échantillon aléatoire simple sans remise, c'est-à-dire, dans notre notation, lorsque k/NM est très petit de façon que $1/(NM - k + 1)$ soit à peu près égal à $1/NM$. Tout semble donc être intuitivement uniforme pour l'ensemble des cas étudiés.

Toutefois, le fait de généraliser en fonction de tailles de grappes inégales M_i en choisissant m_i sous forme de

$$\Pr(m_1 = i_1, \ldots, m_k = i_k) = \frac{\binom{M_1}{i_1} \cdots \binom{M_k}{i_k}}{\binom{\sum_1^N M_i}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \tag{9}$$

n'entraîne pas une distribution de probabilités valable. Soit dit en passant, nous supposerons de nouveau que les grappes originales subissent un échantillonnage à probabilités égales et sans remise, comme à la sous-section 2.4.1. Ensuite (sous-section 2.4.3), nous examinerons des échantillons originaux qui se prêtent à une forme quelconque de sélection PPT (probabilité proportionnelle à la taille).

Pour constater qu'il n'est pas facile de simplement généraliser l'équation (6) sous la forme de l'équation (9), considérons le contre-exemple suivant où la «probabilité» calculée à l'aide de (9) est supérieure à un. Supposons $N = 4$ et des tailles de grappes $M_1 = 4$, $M_2 = 6$, $M_3 = 8$, et $M_4 = 10$. Supposons également que nous prélevons un échantillon en grappes avec $k = 2$ et que, simplement par hasard, les deux grappes retenues sont les plus grandes, c'est-à-dire $M_3 = 8$ et $M_4 = 10$. Il est évident qu'avec ces sélections, l'équation (9) permettrait de produire une probabilité de sélection d'une unité dans chaque grappe qui serait supérieure à un.

Cette difficulté peut-elle être surmontée? Oui, bien que non pas, sans doute, de façon tout à fait satisfaisante. Une méthode consiste à employer une distribution hypergéométrique qui suppose des grappes qui sont toutes aussi grandes que la plus grande grappe de la population. L'inconvénient, c'est que la taille de l'échantillon inverse obtenue n'est plus fixe, et que le sous-échantillon qui en résulte n'est un EAS que conditionnellement, compte tenu de la taille de l'échantillon atteinte, désignée par exemple k_0. Autrement dit, pour un échantillon donné de taille k_0, $k_0 \leq k$, tous les échantillons de taille k_0 ont les mêmes chances d'être choisis à l'aide de l'algorithme inverse.

Soit M_* la taille maximale des grappes, $M_* = \mathrm{Max}\{M_1, M_2, \ldots, M_N\}$. Il s'agit de créer une population en remplissant chacune des grappes originales d'unités «fictives» ou de paramètres substituables, $j = M_i + 1, M_i + 2, \ldots, M_*$. Dès lors, grâce à une méthode semblable à celle de Lahiri (1951) pour l'échantillonnage PPT, l'algorithme inverse permet de choisir des unités de la population constituées de N grappes de taille M_* chacune, puis d'écarter tout élément qui ne se trouve pas dans la «sous-population» constituée des grappes originales de taille M_i.

Plus particulièrement, étant donné un échantillon en grappes constitué de k grappes, il s'agit de choisir le vecteur $\boldsymbol{m}$ de la distribution de probabilités

$$\Pr(m_1 = i_1, \ldots, m_k = i_k) =$$

$$\frac{\binom{M_*}{i_1}\binom{M_*}{i_2} \cdots \binom{M_*}{i_k}}{\binom{NM_*}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \tag{10}$$

où la somme des constituants de $\boldsymbol{m}$ est k et q des constituants m_i ne sont pas zéro. Nous avons là une distribution de probabilités convenable. Étant donné la valeur choisie de m_i, il s'agit de choisir un échantillon aléatoire de m_i unités de la grappe i, où la grappe contient M_i unités de la population et $M_* - M_i$ «paramètres substituables». On écarte toutes les unités choisies qui sont des paramètres substituables, dans l'ensemble de $j = M_i + 1, M_i + 2, \ldots, M_*$. Par conséquent, la taille finale de l'échantillon n'est pas nécessairement égale à k, mais peut lui être inférieure, par exemple k_0.

L'échantillon qui en résulte est conditionnellement un EAS de la population, en ce sens que pour une valeur donnée de k_0, tous les échantillons de taille k_0 ont les mêmes chances d'être choisis à l'aide de cet algorithme inverse. Pour le constater, il suffit de continuer à considérer le problème comme un cas de sous-population, P, de N grappes de taille M_i, $i = 1, \ldots, N$, à l'intérieur d'une population P_* de N grappes, chacune de taille M_*. À noter que, pour tout échantillon, S_*, de taille k tiré de la population P_*, la probabilité de sélection de S_* à l'aide de l'algorithme inverse est la suivante

$$\frac{1}{\binom{NM_*}{k}}. \tag{11}$$

Si $k_0 = k$, c'est là la probabilité de sélection de cet échantillon à l'aide de l'algorithme inverse. Pour une valeur $k_0 < k$, fixe, notons S_0 tout échantillon donné de taille k_0 contenu dans P. Nous pouvons produire un échantillon S_* contenant S_0 en commençant par S_0 et en y ajoutant $k - k_0$ éléments des $N^* M_* - M_+$ paramètres substituables dans P_*. Le nombre d'échantillons S_* de ce type, entraînant une sélection de S_0, est le suivant:

$$\binom{NM_* - M_+}{k - k_0} \quad \text{où} \quad M_+ = \sum_{i=1}^{N} M_i. \tag{12}$$

Par conséquent, la probabilité de sélection de S_0 à l'aide de l'algorithme inverse est égale à la probabilité de sélection de S_* à l'aide de l'algorithme inverse, donnée dans (11), avec sommation sur tous les échantillons S_* construits selon la description ci-dessus, où le nombre d'échantillons de ce genre est donné par (12). Cette probabilité est égale à:

$$\Pr(m_1 = i_1, \ldots, m_k = i_k) = \frac{\binom{M}{i_1} \cdots \binom{M}{i_k}}{\binom{NM}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)} \quad (6)$$

où $0 \le i_j \le k$, $i_1 + i_2 + \ldots + i_k = k$, et q est le nombre de i_j non zéro. Ainsi, avec $M = 100$, $N = 6$, $k = 3$ on a

$$\Pr(m_1 = 1, m_2 = 0, m_3 = 2) = \frac{\binom{100}{1}\binom{100}{0}\binom{100}{2}}{\binom{600}{3}} * \frac{6*5}{3*2}$$

$$\Pr(m_1 = 3, m_2 = 0, m_3 = 0) = \frac{\binom{100}{3}}{\binom{600}{3}} * \frac{6}{3}.$$

Une fois les m_i déterminés, un échantillon aléatoire simple de taille m_i est choisi dans la grappe $i, i = 1, 2, \ldots, k$. Par conséquent, la probabilité conditionnelle de sélection de S_k est

$$\Pr(\text{sélection } S_k \mid S_D) = \frac{1}{\binom{NM}{k}} * \frac{N(N-1)\ldots(N-q+1)}{k(k-1)\ldots(k-q+1)}. \quad (7)$$

La probabilité de sélection d'un échantillon S_k particulier est obtenue en multipliant l'équation (5) par l'équation (7). Il est facile de vérifier que cette opération donne la probabilité correcte de sélection d'un EAS.

Contrairement à l'exemple stratifié, où la fonction de sélection des valeurs de m_i était une fonction de probabilité connue, il n'est pas immédiatement évident que l'équation (6) décrit une distribution de probabilités. Puisque les valeurs établies à l'aide de cette fonction sont toutes non négatives, il suffit de montrer que leur somme égale un sur l'ensemble des valeurs possibles. Le premier facteur de l'équation se présente sous forme d'une distribution hypergéométrique, sauf que le numérateur se limite à k parmi les N grappes, tandis que le dénominateur reflète les N grappes totales. Il est utile de définir une partition de k comme combinaison de nombres entiers positifs dont l'addition donne k, sans égard à l'ordre. Ainsi, les partitions de $k = 3$ sont $\{3\}$, $\{1,2\}$ et $\{1,1,1\}$. Puisque les grappes sont toutes de la même taille, M, tous les motifs de sélection qui correspondent à la même partition sont également probables. Prenons, à titre d'exemple, $N = 6$ et $k = 3$. Dans la pleine distribution hypergéométrique, à taille de grappes égale, chacune des combinaisons ci-dessous a les mêmes chances de se produire:

(0,0,0,0,1,2), (0,0,0,0,2,1), (0,0,0,1,2,0), ..., (2,1,0,0,0,0).

Le nombre total de combinaisons de ce genre est $N(N-1)\ldots(N-q+1)$, où q est la taille de la partition, c'est-à-dire le nombre de valeurs (non zéro) dans la partition.

Dans l'exemple ci-dessus, $q = 2$. Pour une partition donnée, si le nombre de valeurs non zéro peut seulement être placé dans k cellules particulières, on a dès lors $k(k-1)\ldots(k-q+1)$ ordonnancements de ce genre. Par conséquent, la sommation de la distribution sur toutes les valeurs de $(i_1, \ldots, i_k)$ peut se faire en prenant d'abord la somme sur toutes les partitions de k et ensuite pour chaque partition, quitte à prendre la somme sur tous les ordonnancements possibles de cette partition en k cellules. Puisque tous les ordonnancements associés à une partition particulière ont les mêmes chances de se produire, il en résulte une sommation qui correspond à la sommation de la distribution hypergéométrique sur l'espace correct et, par conséquent, la somme de l'expression (6) donne un.

La distribution de probabilités requise pour ce plan de sondage simple en grappes (équation 6) est appréciablement plus difficile à produire que la distribution hypergéométrique dans le cas de l'échantillon stratifié. Toutefois, à mesure que la fraction de sondage k/N diminue, la probabilité est souvent contenue dans deux des partitions seulement: $q = k$ et $q = k - 1$. (Ces probabilités sont calculées à l'annexe.) La probabilité peut même être concentrée uniquement dans le motif avec $q = k$ (on trouvera également à l'annexe un cas spécial de ce phénomène).

Compte tenu des résultats de l'annexe, il est peut-être possible de se rapprocher de l'inverse exact en choisissant un cas dans chaque grappe, en faisant appel à un sondage systématique dans l'échantillon en grappes original. Cette stratégie a une valeur réelle car les calculs de distribution de probabilités deviennent peu maniables à mesure que le nombre de grappes dans l'échantillon augmente. Pour qu'un inverse systématique réussisse, toutefois, l'«étape» doit bien sûr être au moins aussi grande que M sinon plus grande, suivant le nombre de grappes dans la population. Si l'on voulait répéter ce genre de sous-échantillonnage pour chaque inverse systématique d'échantillon, les unités à l'intérieur de chaque grappe seraient ordonnées de nouveau de façon aléatoire avant la prochaine sélection et les grappes triées de nouveau, également de façon aléatoire, après quoi on aurait un autre départ aléatoire avant de passer de nouveau à travers l'échantillon original.

2.4.2 Échantillonnage en grappes à un degré à grappes inégales et à probabilités égales

L'algorithme de sondage inverse pour un échantillon de grappes de taille égale ne se laisse pas généraliser aisément lorsqu'on prélève un échantillon de grappes de taille inégale. Il en est ainsi même s'il semble facile de généraliser cette stratégie d'une façon évidente. En particulier, il ne semble pas difficile de généraliser la méthode précédente de façon que l'on puisse multiplier les «probabilités» et produire la probabilité de sélection «correcte», c'est-à-dire

$$\frac{1}{\binom{M_+}{k}}, \quad \text{où} \quad M_+ = \sum_1^N M_i. \quad (8)$$

de sondage les plus courants: stratifié, en grappes, à degrés multiples et à degrés multiples stratifiés. On y trouve également un exemple d'un algorithme inverse qui à prime abord ne semble pas possible.

2.3 Inversion d'un échantillon stratifié

La présente sous-section introduit un algorithme inverse pour un échantillon stratifié à quatre strates. L'algorithme permet de généraliser pour tout nombre de strates. Nous avons un échantillon stratifié ayant des tailles d'échantillon fixes n_h dans chaque strate h, et des tailles de population de strate connues, $N_1 + N_2 + N_3 + N_4 = N$. Puisqu'un échantillon donné de taille arbitraire m de la population risque d'être contenu entièrement dans une même strate, l'échantillon aléatoire simple le plus grand qui puisse être tiré d'un échantillon stratifié est de taille $m = \min\{n_h\}$.

Pour un échantillon donné S_m, notons (x_1, x_2, x_3, x_4) le nombre d'unités dans chaque strate. Chaque x_i se situe entre 0 et m et $x_1 + x_2 + x_3 + x_4 = m$. La probabilité que S_m soit contenu dans l'échantillon stratifié est égale au nombre d'échantillons stratifiés qui contiennent ces unités m divisé par le nombre total d'échantillons stratifiés possibles, c'est-à-dire

$$\Pr(S_m \subset S_D) = \frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}. \quad (2)$$

L'algorithme qui permet de tirer un EAS de l'échantillon stratifié comporte les trois étapes que voici:

(1) Déterminer la taille des EAS qu'il s'agit de choisir: $m \le \min\{n_h\}$.

(2) Établir une réalisation $\{m_1, ..., m_4\}$ à partir d'une distribution hypergéométrique, à probabilités

$$\Pr(m_1 = i_1, m_2 = i_2, ..., m_4 = i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}} \quad (3)$$

où $i_1 + i_2 + i_3 + i_4 = m$ et $0 \le i_1 \le m$, $0 \le i_2 \le m$, $0 \le i_3 \le m$, $0 \le i_4 \le m$.

(3) Dans chaque strate h, choisir un échantillon aléatoire simple de taille m_h, sans remise, à partir des unités d'échantillonnage n_h.

La probabilité conditionnelle de sélection de l'échantillon S_m, étant donné qu'il est contenu dans l'échantillon stratifié, est donc

$$\frac{\binom{N_1}{x_1}\cdots\binom{N_4}{x_4}}{\binom{N}{m}} \frac{1}{\binom{n_1}{x_1}\cdots\binom{n_4}{x_4}}. \quad (4)$$

La probabilité de sélection d'un échantillon S_m donné quelconque à l'aide de l'algorithme inverse est le produit des deux probabilités données dans les équations (2) et (4). On peut montrer aisément que ce produit est égal à

$$\frac{1}{\binom{N}{m}}.$$

Par conséquent, cette procédure reproduit inconditionnellement un mécanisme d'échantillonnage aléatoire simple, c.-à-d. lorsqu'il est pris sur tous les échantillons stratifiés possibles. À noter que si l'on veut tirer tous les EAS possibles de cette population, il faut répéter toute la séquence à partir du choix d'un échantillon stratifié et poursuivant la démarche jusqu'aux étapes 1 - 3.

2.4 Inversion d'un échantillon en grappes à un degré

Dans la présente sous-section, nous envisageons trois cas spéciaux. Pour commencer, nous examinons des échantillons en grappes dont les grappes sont de taille égale. Nous abordons ensuite le cas plus fréquent où les grappes sont de taille inégale. Dans l'un et l'autre contextes, nous supposons que les grappes sont prélevées à l'aide d'un mécanisme d'échantillonnage aléatoire simple et sans remise. Le troisième cas étudié est celui de grappes inégales échantillonnées à l'aide d'un mécanisme de probabilité proportionnelle à la taille (PPT). Dans ce dernier cas, nous supposons un échantillonnage avec remise.

2.4.1 Échantillonnage en grappes à un degré à tailles de grappes égales et à probabilités égales

Supposons que nous ayons une population de N grappes où toutes les grappes sont de taille M, k d'entre elles étant choisies par un mécanisme d'échantillonnage aléatoire simple sans remise.

Si l'on veut construire un algorithme inverse, il faut décider ce que pourrait être le plus grand sous-échantillon d'éléments. Il est évident que le plus grand EAS d'éléments que l'on peut choisir est k. Soit dit en passant, la taille des grappes n'est pas une contrainte pour ce qui est de la taille du sous-échantillon.

Pour un échantillon S_k donné, notons q le nombre de grappes représenté dans S_k; $0 < q \le k$. Dès lors la probabilité que S_k soit contenu dans l'échantillon en grappes est égale au nombre d'échantillons en grappes qui contiennent ces grappes q divisé par le nombre total d'échantillons en grappes possibles, c'est-à-dire

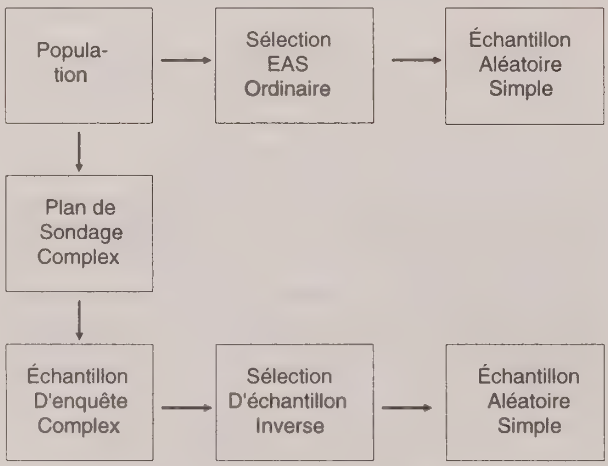
$$\Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}. \quad (5)$$

Comme pour l'échantillon stratifié, l'algorithme détermine d'abord le nombre d'unités à choisir dans chaque grappe, $(m_1, m_2, ..., m_k)$. La distribution de probabilités à utiliser pour choisir les m_i est

Nous sommes d'avis qu'il y a parfois lieu de répondre à cette question par un «Oui». Nous appelons ce deuxième plan de sondage un «algorithme de plan de sondage inverse», d'où le titre du présent exposé.

Un schéma peut nous aider à visualiser l'algorithme (voir la figure 1). Dans le diagramme, deux stratégies d'échantillonnage sont comparées, toutes deux donnant des échantillons aléatoires simples d'une population:

(1) Le premier plan (rangée du haut) fait appel à un processus conventionnel de sélection directe d'échantillons aléatoires simples (EAS) (p. ex., Cochran 1977), de sorte que tous les échantillons possibles d'une taille donnée comportent la même probabilité de sélection. (Ce type de plan est souvent peu commode ou inefficace ou les deux à la fois, et c'est pourquoi on ne les utilise à peu près jamais, même s'il en est question dans les manuels.)
(2) Le deuxième plan représente un processus en deux étapes. La première étape consiste à échantillonner la population d'une façon complexe qui s'appuie soigneusement sur la nature de la population et des besoins de la clientèle, les ressources de la clientèle étant utilisées sobrement (c'est là le domaine par excellence des concepteurs d'enquête).
(3) Notre formulation a ceci de nouveau qu'elle prélève un second échantillon (peut-être complexe?) qui inverse le premier ensemble de sélections de façon à fournir à terme un échantillon aléatoire simple. Bien entendu, il serait inefficace d'utiliser ce processus en deux étapes pour tirer un échantillon aléatoire simple unique d'une enquête complexe normalement beaucoup plus grande, et c'est pourquoi nous proposons la création d'échantillons aléatoires simples multiples, nos inférences étant fondées sur ceux-ci.



Bien qu'il soit possible d'élaborer, la nature fondamentale de algorithmes dont nous parlons devrait maintenant être évidente. Ils peuvent ne comporter que quatre étapes de base:

(1) Inverser, dans la mesure du possible, le plan complexe existant de façon à pouvoir créer des sous-échantillons aléatoires simples (à un degré d'approximation utile).
(2) Possiblement, appliquer le logiciel statistique conventionnel directement au sous-échantillon, puisque cela est désormais approprié.
(3) Répéter le sous-échantillonnage et l'analyse conventionnelle des étapes (1) et (2) de façon répétée.
(4) Retenir, dans la mesure du possible, le caractère du paradigme de randomisation original en utilisant la distribution des résultats du sous-échantillon comme base d'inférence (au lieu de l'échantillon complexe original).

Soulignons ce que cette stratégie est et n'est pas: tout d'abord, elle est très exigeante en traitement informatique, se fondant sur des calculs bon marché et même très bon marché. Deuxièmement, elle suppose l'existence d'algorithmes inverses pratiques (ce qui n'est pas toujours le cas). Troisièmement, elle suppose que la robustesse originale de l'échantillon tout entier peut être captée si l'on prélève un nombre suffisant de sous-échantillons, de façon qu'il n'y ait aucune perte appréciable d'efficacité. Quatrièmement, même si elle ressemble à la méthode bootstrap (Efron 1979), il ne s'agit pas de bootstrap. Il n'y a aucune intention d'imiter les sélections originales, comme il faudrait le faire pour bien utiliser la méthode bootstrap (p. ex. McCarthy et Snowden 1985; Rao et Wu 1988). Tout au contraire; notre but ici est de créer, à partir du plan original, un ensemble de sous-échantillons totalement différent et plus facilement analysable.

2.2 Définition d'un algorithme de sondage inverse

Supposons que nous voulons tirer un échantillon aléatoire simple, sans remise, d'une population finie de taille N. Supposons, également, que la population ne soit plus disponible pour l'échantillonnage, mais que nous possédions un échantillon choisi à même cette population à l'aide d'un plan de sondage D; notons cet échantillon S_D. Soit S_m, un deuxième échantillon de taille m qui pourrait être tiré de la population. Un algorithme de sondage inverse doit décrire la façon de choisir un échantillon de S_D de façon que pour tout échantillon S_m donné

$$\Pr(\text{sélection } S_m \mid S_D) * \Pr(S_m \subset S_D) = \frac{1}{\binom{N}{m}}. \qquad (1)$$

La première étape consiste à calculer la probabilité qu'un échantillon S_m arbitraire mais fixe soit contenu dans l'échantillon S_D. Bien entendu, il existe des contraintes quant à la taille de l'échantillon aléatoire simple (EAS) qui peut être prélevé de cette façon; la probabilité que S_D contienne S_m ne peut pas être zéro. Par conséquent, la taille de l'EAS ne peut certainement pas être supérieure à celle de l'échantillon S_D original et, en effet, la taille de l'EAS doit généralement être bien inférieure à celle de l'échantillon complexe original.

Il s'agit donc de trouver un algorithme général capable de tirer un EAS d'un échantillon S_D donné selon une probabilité conditionnelle correcte. Il est également nécessaire de s'assurer que l'on utilise des fonctions de probabilité valables. Les sous-sections qui suivent illustrent les algorithmes de sondage inverse pour quelques-uns des plans

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 13-24
Statistique Canada

Algorithmes de plan de sondage inverses

SUSAN HINKINS, H. LOCK OH, et FRITZ SCHEUREN[1]

RÉSUMÉ

Dans le travail ordinaire en statistique, l'échantillonnage est souvent exécuté en fonction d'un processus qui choisit des variables aléatoires telles qu'elles sont indépendantes et distribuées de façon identique (IDI). D'importantes techniques comme la régression et l'analyse des tableaux de contingence ont été élaborées largement dans ce contexte IDI, de sorte qu'il faut avoir recours à des rajustements pour les utiliser dans le contexte d'une enquête complexe. Toutefois, au lieu de rajuster l'analyse, les auteurs ont adopté une formulation qui a ceci de nouveau qu'elle prélève un second échantillon dans l'échantillon original. Dans ce second échantillon, le premier ensemble de sélections est inversé de façon à fournir à terme un échantillon aléatoire simple. Bien entendu, il serait inefficace d'utiliser ce processus en deux étapes pour tirer un échantillon aléatoire simple unique d'une enquête complexe normalement beaucoup plus grande, et c'est pourquoi des échantillons aléatoires simples multiples sont prélevés, les auteurs ayant élaboré une façon de fonder sur eux des inférences. Les échantillons originaux ne peuvent pas tous être inversés, mais les auteurs abordent de nombreux cas spéciaux qui couvrent tout un éventail de possibilités.

MOTS CLÉS: Échantillonnage de populations finies; inférence dans les enquêtes complexes; rééchantillonnage.

1. INTRODUCTION

L'évolution des enquêtes par échantillonnage comtemporaines est un phénomène extraordinaire (Bellhouse 1988; Hansen 1987; Kish 1995). La richesse même de cette évolution a peut-être entraîné, par contre, l'isolement des enquêtes par échantillonnage du reste du secteur statistique, l'attention y étant accordée à la richesse des modèles. En effet, il est bien connu que, dans le travail ordinaire en statistique, l'échantillonnage est souvent exécuté en fonction d'un processus qui choisit des variables aléatoires telles qu'elles sont indépendantes et distribuées de façon identique (IDI).

D'importantes techniques comme la régression et l'analyse des tableaux de contingences ont été élaborées largement dans ce contexte IDI, de sorte qu'il faut avoir recours à des rajustements pour les utiliser dans le contexte d'une enquête complexe. Des ouvrages entiers ont même été consacrés à cette question (Skinner, Holt et Smith 1989) et on y a accordé beaucoup de temps et d'efforts dans des logiciels (comme SUDAAN ou WESVAR PC) préparés expressément pour les enquêtes (voir aussi Wolter 1985). Compte tenu de tout ce qui a été accompli déjà, y a-t-il quelque chose de valable à ajouter? Nous proposons une façon de mieux traiter l'«interface» que l'on trouve actuellement entre l'IDI et la statistique des enquêtes.

Le présent exposé est divisé en quatre sections. Cette introduction est la section 1. Dans les sections 2 et 3 on trouve un énoncé général du problème et plusieurs «résolutions» pour quelques-uns des plans les mieux connus. Notre stratégie consiste à rééchantillonner l'échantillon complexe de façon à obtenir une structure des données plus facile à analyser. En particulier, nous abordons l'échantillonnage d'éléments stratifiés, les échantillons en grappes à un et deux degrés, ainsi que le plan important de deux unités primaires d'échantillonnage par strate (section 2). Puisqu'il est peu probable qu'un rééchantillonnage donné contienne toute l'information de l'enquête originale, nous vérifions ce qui se produit lorsque l'échantillon complexe original est rééchantillonné de façon répétée. On trouve également à la section 3 une illustration concrète de nos idées, tirée de notre pratique et fondée sur un échantillon SDR (statistique des revenus) hautement stratifiée de déclarations de revenus des sociétés (p. ex. Hughes, Mulrow, Hinkins, Collins et Uberall 1994). Enfin, à la section 4, nous décrivons quelques applications et les prochaines étapes qui rendront nos idées embryonnaires encore plus utiles.

2. ÉNONCÉ DU PROBLÈME ET «RÉSOLUTIONS» POSSIBLES

2.1 Raisonnement et stratégie de base

À supposer que nous voulions appliquer une procédure IDI à un échantillon d'enquête complexe. À supposer, également, que nous voulions jeter un regard neuf sur la façon de «résoudre» le problème d'interface qui survient parce que le plan d'enquête n'est pas du type IDI. Comment procéder? Il existe une expression courante qui semble résumer notre stratégie:

Si vous n'avez qu'un marteau, tout problème devient un clou.

À titre d'échantillonneurs, nous avons un marteau et c'est le processus d'échantillonnage. Pouvons-nous transformer le problème d'interface des enquêtes en un clou dont nous pourrions nous occuper à l'aide d'un autre plan de sondage?

[1] Susan Hinkins, Internal Revenue Service, Bozman, MT, U.S.A.; H. Lock Oh, Internal Revenue Service, Washington, DC, U.S.A.; Fritz Scheuren, Ernest and Young, 1402 Ruffner Rd., Alexandria, VA 22302 U.S.A.

NATH, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

NATH, S.N. (1969). More results on product moments from a finite universe. *Journal of the American Statistical Association*, 64, 864-869.

RAGHUNANDANAN, K., et SRINIVASAN, R. (1973). Some product moments useful in sampling theory. *Journal of the American Statistical Association*, 68, 409-413.

STAFFORD, J.E. (1996). A note on symbolic Newton-Raphson, submitted for publication.

STAFFORD, J.E., et ANDREWS, D.F. (1993). A symbolic algorithm for studying adjustments to the profile likelihood. *Biometrika*, 80, 715-730.

WISHART, J. (1952). Moment coefficients of the *k*-statistics in samples from a finite population. *Biometrika*, 39, 1-13.

illustrer avec l'estimateur de Horvitz-Thompson de $\sum\sum$ donné par $(n/N)m(y/\pi)$ dans la notation élaborée ici. L'application de l'opérateur $Cum[\cdot]$ dans le cadre d'un plan d'échantillonnage général en vue de l'obtention du troisième cumulant de l'estimateur de Horvitz-Thompson donne

$$2\frac{\left\{\sum_{i=1}^{N} y_i\right\}^3}{N^3} - 3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\frac{y_i^2}{\pi_i}\right\}}{N^3} - 3\frac{\sum_{i=1}^{N}\frac{y_i^3}{\pi_i^2}}{N^3}$$

$$-3\frac{\left\{\sum_{i=1}^{N} y_i\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}{}'\frac{\pi_{ij}y_iy_j}{(\pi_i\pi_j)}\right\}}{N^3} + 3\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}y_iy_j^2}{(\pi_i\pi_j^2)}}{N^3}$$

$$+\sum_{i=1}^{N}\sum_{h=1}^{N}\sum_{k=1}^{N}\frac{\frac{\pi_{ijk}y_iy_jy_k}{(\pi_i\pi_j\pi_k)}}{N^3}$$

où, par exemple, le terme π_{ii} est la probabilité d'inclusion simple π_i.

L'opérateur $Aexp[\cdot]$ a deux arguments, c'est-à-dire la fonction pour laquelle le développement est requis et l'ordre du développement. Cet opérateur est utilisé avec les opérateurs $EV[\cdot]$ ou $Cum[\cdot]$ en vue de l'obtention de cumulants ou espérances approximatives. On peut l'illustrer dans le cas de l'estimateur de régression linéaire multiple dans le cadre d'un sondage aléatoire simple sans remise. Lorsqu'il y a q covariables l'estimateur de régression résultant est donné par

$$k(y) + b_{i_1}[K(x^{i_1}) - k(x^{i_1})] \tag{29}$$

à l'aide de la notation des indices et des statistiques k. Dans (29) le coefficient b_{i_1} est le vecteur qui résulte du produit $k(x_{i_1}, y)\,ik(x^{i_1}, x_{i_2})$ dans la notation des indices, où le tableau $q \times q$ $ik(x_{i_1}, x_{i_2})$ est l'inverse du tableau $q \times q$ donné par $k(x_{i_1}, x_{i_2})$. De même, nous pouvons utiliser $IK(x_{i_1}, x_{i_2})$ pour désigner l'inverse du tableau de population finie $K(x_{i_1}, x_{i_2})$. La dérivation de l'erreur quadratique moyenne de (29) suppose des développements en séries de Taylor des éléments de b_{i_1} suivies des calculs de moments et de la collecte de termes appropriés. La commande *Mathematica* qui permet d'obtenir la variance approximative de (29) est obtenue en appliquant d'abord $Aexp[\cdot]$ à (29) avec 2 pour l'ordre du développement. L'opérateur $Cum[\cdot]$ est ensuite appliqué aux résultats avec les arguments suivants: le résultat du développement asymptotique comme estimateur, le sondage aléatoire simple comme plan et 2 pour l'ordre du cumulant. Cela donne

$$\frac{(N-n)K(y,y)}{Nn} + \frac{(-N+n)K(x_{i_1},y)K(x_{i_2},y)IK(x^{i_1},x^{i_2})}{Nn}$$

dans la notation des indices comme sortie.

L'estimation est réalisée par l'entremise de l'opérateur $UE[\cdot]$ qui comporte deux arguments, l'estimande et le plan d'échantillonnage. Ainsi, l'application de $UE[\cdot]$ à $\{M(x)\}^2$ dans le cadre d'un sondage aléatoire simple donne

$$\frac{(Nn)\{k(x)\}^2 + (N-n)k(x,x)}{Nn}.$$

Si l'estimande ne se laisse pas exprimer comme une somme de sommes imbriquées, mais se laisse plutôt exprimer comme la racine d'une fonction estimative, $UE[\cdot]$ donne un estimateur convergent.

7. DISCUSSION AU SUJET DES TRAVAUX À VENIR

Les éléments de base de l'élaboration d'une algèbre informatique globale destinée à la théorie des enquêtes par échantillonnage ont été établis. Cette algèbre se fonde sur l'énumération de partitions. Les opérations de base dans le cadre de l'énumération des partitions englobent l'évaluation de sommes imbriquées et des développements en séries de Taylor. Ces opérations terminées, on peut calculer les espérances de statistiques d'échantillonnage ou déterminer les estimateurs non biaisés de quantités de population.

La prochaine étape du travail consiste à étendre les résultats à un degré à des sondages à plusieurs degrés et à plusieurs phases. Tant pour les sondages à plusieurs degrés que pour les sondages à plusieurs phases, le problème se réduit à une évaluation informatique de sommes multiples dans le cadre d'un opérateur d'espérance ou à la détermination d'un estimateur non biaisé de sommes multiples de population finie. Le problème des sondages à plusieurs degrés est actuellement à l'étude. On cherche également à étendre l'algèbre à des modèles de super-population.

Lorsque l'algèbre de base aura été mise en place, on pourra aisément étudier des problèmes de recherche mettant en cause des formules d'échantillonnage complexes du point de vue algébrique.

REMERCIEMENTS

Les auteurs tiennent à remercier David Andrews de ses discussions utiles. Les auteurs ont bénéficié de subventions du Conseil de recherches en sciences naturelles et en génie du Canada et d'un contrat de recherches de Statistique Canada.

BIBLIOGRAPHIE

ANDREWS, D.F., et STAFFORD, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society (B),* 55: 613-628.

KENDALL, W.S. (1993). Computer algebra in probability and statistics. *Statistica Neerlandica,* 47, 9-25.

McCULLAGH, P. (1987). *Tensor Methods in Statistics.* New York: Chapman and Hall.

(1,1,2). Si l'on soustrait 1 de chaque valeur d'indice dans la liste, on obtient la liste (1,0,0), (0,1,0), (0,0,1). Par conséquent le terme requis dans le développement est $(e_{11}e_{02}e_{03}+e_{01}e_{12}e_{03}+e_{01}e_{02}e_{13})/\sqrt{n}$ ou par équivalence $[z(m(y))-M(y)z(m(x))/M(x)]/\sqrt{n}$. Le terme $1/n$ est obtenu de (24) qui se réduit à

$$[M(y)\{z(x)\}^2/\{M(x)\}^2 - z(x)z(y)/M(x)]/n.$$

L'estimateur de régression qui se trouve dans (26) peut être exprimé sous la forme

$$K(y) + \frac{z(k(y))}{\sqrt{n}} + \left[K(x,y) + \frac{z(k(x,y))}{\sqrt{n}}\right] \times \left[K(x,x) + \frac{z(k(x,x))}{\sqrt{n}}\right]^{-1}\left[\frac{z(k(x))}{\sqrt{n}}\right] \qquad (28)$$

à l'aide de (3). Les termes entre crochets dans (28) peuvent donner lieu à un développement semblable à celle qui donne l'estimateur de quotient. Dans ce cas les termes des développements deviennent: $e_{01} = K(x,y)$, $e_{11} = z(k(x,y))$ et $e_{21} = e_{31} = \cdots = 0$; $e_{i2} = (-1)^i\{z(k(x,x))\}^i/\{K(x,x)\}^{i+1}$ pour $i = 0, 1, 2, \ldots$; et $e_{03} = 0$, $e_{13} = z(k(x))$ et $e_{23} = e_{33} = \cdots = 0$. Par conséquent, le terme $1/\sqrt{n}$ du développement des termes entre crochets dans (28) est

$$-\frac{K(x,y)z(k(x))}{K(x,x)\sqrt{n}}$$

et le terme $1/n$ est

$$-\frac{1}{n}\left[\frac{z(k(x,y))}{K(x,x)} - \frac{K(x,y)z(k(x,x))}{K(x,x)^2}\right]z(k(x)).$$

Ceux-ci ont été obtenus à l'aide du même argument utilisé dans l'estimateur de quotient.

6. APPLICATIONS MACHINE AU CALCUL DES VALEURS ESPÉRÉES D'UNE VARIABLE STATISTIQUE ET À L'OBTENTION D'ESTIMATEURS NON BIAISÉS

Puisque l'application machine à la méthode décrite dans les sections 3 à 5 a été effectuée en langage de programmation *Mathematica*, nous présentons une brève description de l'utilisation de *Mathematica*. Nous décrivons ensuite les opérateurs qui ont été élaborés dans *Mathematica* afin de présenter une algèbre informatique destinée à la théorie des enquêtes par échantillonnage.

La programmation *Mathematica* utilise des expressions du type $h[e_1, e_2, \ldots]$ où l'objet h est la tête de l'expression et les e sont les éléments de l'expression. Nous avons élaboré un certain nombre d'expressions machine dans *Mathematica* de type $h[e_1, e_2, \ldots]$ pour des opérateurs que nous appliquons à la mise au point d'une algèbre informatique pour l'échantillonnage. Tous ces opérateurs ont été conçus de façon que leurs arguments puissent prendre la forme de vecteurs aussi bien que de grandeurs scalaires. Il existe quatre opérateurs de base: $EV[\cdot]$ pour la valeur espérée, $Cum[\cdot]$ pour le calcul des cumulants, $UE[\cdot]$ pour l'estimateur non biaisé et $Aexp[\cdot]$ pour le développement asymptotique. Il existe également un opérateur qui sert à passer de la notation axée sur les statistiques k à une notation axée sur des moyennes et inversement.

L'opérateur de valeur espérée $EV[\cdot]$ de la statistique d'échantillonnage combine et exécute dans *Mathematica* les trois opérations de base indiquées dans le schéma de (10). $EV[\cdot]$ contient deux arguments, le premier étant l'expression pour laquelle la valeur espérée doit être obtenue et le deuxième le plan d'échantillonnage qui définit les probabilités d'inclusion. L'application dans *Mathematica* de $EV[\cdot]$ à $m(x_{i_1})m(x_{i_2})m(x_{i_3})$ dans le cadre d'un sondage aléatoire simple sans remise donne

$$K(x_{i_1})K(x_{i_2})K(x_{i_3}) + \frac{(N-n)(K(x_{i_1},x_{i_2})K(x_{i_3})}{Nn} + \frac{K(x_{i_1},x_{i_3})K(x_{i_2}) + K(x_{i_1})K(X_{i_2},x_{i_3}))}{Nn} + \frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2}$$

dans l'expression la plus simple de la sortie. À noter que le résultat est une fonction de la partition complète de $\{i_1, i_2, i_3\}$. Si l'opérande est changé à $\{m(x_{i_1}) - M(x_{i_1})\} \times \{m(x_{i_2}) - M(x_{i_2})\} \times \{m(x_{i_3}) - M(x_{i_3})\}$, l'application de $EV[\cdot]$ donne

$$\frac{(N^2-3Nn+2n^2)K(x_{i_1},x_{i_2},x_{i_3})}{N^2n^2},$$

que Nath (1968) a obtenu pour des valeurs particulières des indices i_1, i_2 et i_3. En réalité, les résultats que l'on trouve dans Nath (1968, 1969) pour les produits de trois et quatre moyennes et les résultats exacts que l'on trouve dans Raghunandanan et Srinivasan (1973) jusqu'à un produit de huit moyennes peuvent tous être reproduits automatiquement à l'aide du logiciel qui a été mis au point.

Jusqu'à présent, le plan d'échantillonnage utilisé dans chacun des exemples a été un sondage aléatoire simple sans remise. Il est possible d'obtenir des résultats dans le cadre de plans d'échantillonnage généraux. Nous illustrons ces résultats pour l'opérateur $Cum[\cdot]$ qui sert à obtenir les cumulants d'un estimateur. À noter que le deuxième cumulant pour un estimateur est également la variance. L'opérateur $Cum[\cdot]$ a trois arguments. Le premier est une expression de l'estimateur, le deuxième est l'ordre du cumulant et le troisième est le plan d'échantillonnage. Dans le cadre de plans d'échantillonnage généraux les estimateurs peuvent être exprimés sous la forme $\sum\prod$ dans le schéma donné par (10) et on peut procéder au développement de $\sum\prod$ pour obtenir $\sum\sum$, le terme moyen dans (10). Toutefois, il n'existe pas de simplification générale pour l'obtention du terme final dans (10), ce qui se laisse

partie inclusion de la règle. L'application répétée de (21) donne

$$\sum_{j_1 \neq \dots \neq j_t = 1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r} \right) = \sum_{P_t \in \wp_t} (-1)^{|J_t| - |P_t|} \times \left\{ \prod_{b_k \in P_t} \left[(|b_k| - 1)! \sum_{j=1}^{N} \left(\prod_{i_r \in b_k} x_{i_r j} \right) \right] \right\}$$

où $|J_t|$, $|P_t|$ et $|b_k|$ sont le nombre d'indices dans J_t, le nombre de blocs dans la partition simple P_t et le nombre d'éléments dans le bloc b_k respectivement.

5. PARTITIONS DE NOMBRES ENTIERS ET LINÉARISATION DE TAYLOR

Supposons que dans le cadre d'un plan d'échantillonnage quelconque un estimateur $\hat{\theta}$ d'un paramètre θ présente un certain intérêt. La méthode décrite dans les sections 2 à 4 peut être utilisée pour les calculs de moments pour $\hat{\theta}$ ou pour trouver des estimateurs non biaisés de ces moments. Ce n'est que dans les cas les plus simples que cette méthode peut être appliquée directement. Typiquement $\hat{\theta}$ doit être linéarisé de façon à devenir une fonction polynomiale de moyennes ou de sommes d'échantillons qui sont des variables aléatoires $O_p(1)$ relativement au plan d'échantillonnage. Lorsque $\hat{\theta}$ a été linéarisé de cette façon, la méthode des sections 2 à 4 est applicable.

L'objectif de la linéarisation est d'écrire $\hat{\theta}$ sous forme de développement asymptotique où les termes se suivent dans l'ordre descendant selon $1/\sqrt{n}$, plus particulièrement

$$\hat{\theta} = \hat{\theta}_0 + \hat{\theta}_1/\sqrt{n} + \hat{\theta}_2/n + \dots, \tag{22}$$

où $\hat{\theta}_i$ est le coefficient du $n^{-i/2}$ terme. Typiquement $\hat{\theta}$ est un produit de quantités qui se prête également à ce type de développement. Par exemple, si la mesure d'intérêt est y et si une variable auxiliaire x est présente, θ pourrait dès lors être $M(y)$ et l'information auxiliaire disponible est $M(x)$ de même que x_j pour $j \in s$. Dès lors $\hat{\theta} = M(x)m(y)/m(x)$, l'estimateur de quotient simple, est un produit de trois quantités $M(x)$, $m(y)$ et $1/m(x)$ ayant toutes des développements asymptotiques individuels. Le développement de $M(x)$ est elle-même. À partir de (3) le développement pour $m(y)$ donne $M(y) + z(m(y))/\sqrt{n}$. Le développement pour $1/m(x)$ est le résultat de (3) puis de l'application d'un développement en séries de Taylor à $[M(x) + z(m(x))/\sqrt{n}]^{-1}$.

En général, on peut trouver toute développement d'une fonction de régularité suffisante si l'on définit des opérateurs pour le développement d'une fonction, par exemple $g(\dot{e})$ où $\dot{e}$ est elle-même un développement. Nous recherchons le développement de fonctions de type

$$g(\dot{e}) = \prod_{j=1}^{p} \dot{e}_j. \tag{23}$$

où $\dot{e}_j$ elle-même comporte le développement $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$. Au moment de linéariser $\hat{\theta}$ l'exigence de base est de définir un opérateur qui reprend $\hat{\theta}_i$ dans (22). L'efficacité de cet opérateur est tirée uniquement d'une règle pour le développement de fonctions du type donné dans (23). Les calculs requis sont des fonctions de partitions de nombres entiers. Par exemple, le terme $1/n$ du développement de $\prod_{j=1}^{3} \dot{e}_j$ est

$$e_{21}e_{02}e_{03} + e_{01}e_{22}e_{03} + e_{01}e_{02}e_{23} + e_{11}e_{12}e_{13} + e_{11}e_{02}e_{13} + e_{01}e_{12}e_{13}. \tag{24}$$

Le fait de rassembler les premiers indices pour chaque terme de la somme permet d'établir une liste dans laquelle chaque élément participe à la somme 2: {(2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)}. En notant que le terme d'ordre $n^{-i/2}$ dans tout développement $\dot{e}_j$ est en réalité le $(i+1)$-ième terme de la somme $\sum_{i=0}^{\infty} e_{ij} n^{-i/2}$, nous pouvons modifier la liste tirée de (24) de façon que les entrées identifient la position des termes dans une somme. La modification consiste à ajouter 1 à chaque valeur d'indice de la liste. Dans la liste tirée de (25), il en résulte un regroupement de toutes les partitions du nombre entier 5 en 3 blocs: {(3,1,1), (1,3,1), (1,1,3), (2,2,1), (2,1,2), (1,2,2)}. En général, l'i-ième terme du développement de $\prod_{j=1}^{p} \dot{e}_j$ ou $\dot{e}_j^{\,p}$, où p est un nombre entier positif, est une somme sur toutes les partitions du nombre entier $i + p$ en p blocs. Par conséquent, si l'on utilise cette méthode, il est possible de trouver tout terme du développement de l'estimateur d'un quotient, par exemple.

Illustrons cette technique à l'aide d'une estimation de quotient et de régression. L'estimateur de quotient est donné par

$$M(x)\,m(y)/m(x) \tag{25}$$

et l'estimateur de régression par

$$k(y) + b[K(x) - k(x)] = k(y) + \frac{k(x,y)}{k(x,x)}[K(x) - k(x)] \tag{26}$$

dans la notation de statistiques k.

Lorsqu'on utilise (3) l'estimateur de quotient (25) peut être exprimé sous la forme

$$M(x)\left[M(y) + \frac{z(y)}{\sqrt{n}}\right]\left[M(x) + \frac{z(x)}{\sqrt{n}}\right]^{-1}. \tag{27}$$

L'expression que l'on trouve dans (27) peut être formulée en fonction de (24) avec $p = 3$. Le premier terme de (27) est le développement $\sum_{i=0}^{\infty} e_{i1} n^{-i/2}$ avec $e_{01} = M(x)$ et $e_{11} = e_{21} = \dots = 0$. Le premier terme entre crochets que l'on trouve dans (28) est le développement $\sum_{i=0}^{\infty} e_{i2} n^{-i/2}$ où $e_{02} = M(y)$, $e_{12} = z(m(y))$ et $e_{22} = e_{32} = \dots = 0$. Le deuxième terme entre crochets est le développement $\sum_{i=0}^{\infty} e_{i3} n^{-i/2}$ où $e_{i3} = (-1)^i \{z(m(y))\}^i / \{M(x)\}^{i+1}$. Afin d'obtenir le terme $1/\sqrt{n}$ dans le développement de (27), dans lequel cas $i = 1$ et $p = 3$, nous devons trouver les partitions de nombres entiers de 4 en blocs de 3. Cela donne les partitions (2,1,1), (1,2,1) et

produit des sommes de façon à identifier les termes pour lesquels l'opérateur d'espérance de la population finie manifestera un comportement différent à cause de différences dans les valeurs des probabilités d'inclusion et des probabilités d'inclusion conjointes.

Par exemple, le produit des sommes $\sum_{j\in s} x_{i_1 j} \sum_{j\in s} x_{i_2 j} \sum_{j\in s} x_{i_3 j}$ peut s'exprimer sous la forme

$$\sum_{j\in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 j} x_{i_3 k} + \sum_{j\neq k\in s} x_{i_1 j} x_{i_2 k} x_{i_3 j} + \sum_{j\neq k\in s} x_{i_1 k} x_{i_2 j} x_{i_3 j} + \sum_{j\neq k\neq l\in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}. \quad (16)$$

Le résultat correspond à la partition complète des indices $I_3 = \{i_1, i_2, i_3\}$ donnés par $\wp_3$ dans (15). L'ordre des partitions dans $\wp_3$ est le même que l'ordre donné pour les termes dans (16). Pour chaque partition dans $\wp_3$, les variables qui se trouvent dans le même bloc comportent le même deuxième indice dans le terme approprié dans (16). Par exemple, la partition $(i_1 i_3 \mid i_2)$ correspond au terme $\sum_{j\neq k\in S} x_{i_1 j} x_{i_2 k} x_{i_3 j}$ dans (16). Chaque terme du résultat peut être identifié par une partition de I_3 et chaque partition détermine le comportement de l'opérateur de la valeur attendue.

En général, nous souhaitons un développement des produits de type $\prod_{r=1}^{m} \sum_{j\in s} x_{i_r j}$, où le produit est pris sur les éléments i_r de l'ensemble d'indices $I_m = \{i_1, ..., i_m\}$. Comme dans (16), le produit peut être exprimé en fonction de la partition complète de I_m. Il en est ainsi parce que la règle itérative qui régit le développement d'un produit de sommes imite la règle d'inclusion-exclusion.

Le développement des produits de sommes par l'entremise de partitions est démontrée par induction comme suit. Supposons que le produit des $t-1$ premières sommes puisse être exprimé comme une somme sur la partition complète de l'ensemble d'indices $I_{t-1} = \{i_1, ..., i_{t-1}\}$, en particulier

$$\prod_{r=1}^{t-1} \left(\sum_{j\in s} x_{i_r j}\right) = \sum_{P_{t-1}\in \wp_{t-1}} X_{P_{t-1}}. \quad (17)$$

Dans (17) le terme $X_{P_{t-1}}$ est la somme identifiée par la partition $P_{t-1} = (b_1 \mid ... \mid b_k), k = 1, ..., t-1$. Les blocs b_j indiquent des groupes de variables ayant le même deuxième indice et donc P_{t-1} produit un ensemble $J_k = \{j_1, ..., j_k\}$ de deuxièmes indices. On peut exprimer $X_{P_{t-1}}$ sous la forme

$$X_{P_{t-1}} = \sum_{j_1\neq ...\neq j_k\in s} \left(\prod_{j\in J_k} X_{b_j}\right), \quad (18)$$

où X_{b_j} est un produit de x définis par le bloc b_j ayant tous le même deuxième indice. À titre d'illustration de (18), considérons, par exemple, le troisième terme de (16). Ici $P_{t-1} = (i_1 i_3 \mid i_2)$ et $J_2 = \{j, k\}$ de sorte que dans (18) la somme est prise sur $j\neq k\in s$ et les multiplicandes du produit sont $X_{b_j} = x_{i_1 j} x_{i_3 j}$ et $X_{b_k} = x_{i_2 k}$. Pour ce qui est de la discussion générale, lorsque l'un ou l'autre des membres de (17) est multiplié par $\sum_{j\in s} x_{i_t j}$ on obtient le produit des t premières sommes. Le produit $X_{P_{t-1}} \sum_{j\in s} x_{i_t j}$ peut donc s'exprimer sous la forme

$$\sum_{j_1\neq ...\neq j_k\in s} \left(\sum_{l=1}^{k} x_{i_t j_l} \prod_{j\in J_k} X_{b_j}\right) + \sum_{j_1\neq ...\neq j_k\neq j_{k+1}\in s} \left(\prod_{j\in J_k} X_{b_j} x_{i_t j_{k+1}}\right). \quad (19)$$

Le premier terme de (19) correspond à la partie inclusion de la règle et le deuxième terme de (19) correspond à la partie exclusion de la règle. Lorsqu'on établit la somme de (19) pour tous les $P_{t-1}\in \wp_{t-1}$, le résultat est une somme pour la partition complète des t premiers indices donnés par I_t, c.-à-d. la somme pour tous les $P_t \in \wp_t$.

Lorsqu'on a procédé au développement du produit des sommes $\prod_{r=1}^{m} \sum_{j\in s} x_{i_r j}$ en une somme de sommes imbriquées, l'opérateur de valeur espérée de la population finie peut être appliqué à chaque terme de façon que l'on obtienne la valeur espérée de ce produit. La valeur attendue dans le cadre d'un sondage aléatoire simple sans remise du produit des sommes donne lieu à une somme pondérée de sommes imbriquées, chaque somme étant prise sur la population finie. Il s'agit ensuite d'évaluer ces sommes imbriquées.

En général, il s'agit d'évaluer la somme imbriquée $\sum_{J_t} Y_{J_t}$ où J_t est l'ensemble d'indices $\{j_1, ..., j_t\}$. La somme est prise sur tous les $j_1 \neq ... \neq j_t$ avec chaque $j_r = 1, ..., N$. L'opérande Y_{J_t} est le produit $x_{i_1 j_1} x_{i_2 j_2} ... x_{i_t j_t}$. Dans le cas spécial où $t = 3$ où $J_3 = \{j, k, l\}$ la somme imbriquée peut s'écrire sous la forme de sommes complètes:

$$\sum_{J_3} Y_{jkl} = \sum_{j\neq k\neq l=1}^{N} Y_{jkl} = \sum_{j\neq k\neq l=1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} =$$

$$2\sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j} -$$

$$\sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} x_{i_3 j} + \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_2 j} \sum_{j=1}^{N} x_{i_3 j}. \quad (20)$$

À noter que les sommes complètes de l'expression à l'extrême droite dans (20) sont le résultat de la partition complète $\wp_3$ dans (15). L'ordre des partitions dans $\wp_3$ est le même que l'ordre des termes à la droite de (20). Les indices inférieurs à la droite de (20) désignent les membres de blocs dans $\wp_3$. Par exemple, la partition $(i_1 i_3 \mid i_2)$ correspond au terme $\sum_{j=1}^{N} x_{i_1 j} x_{i_3 j} \sum_{j=1}^{N} x_{i_2 j}$ dans (20). À noter également dans (20) que la détermination d'une somme imbriquée est rendue complexe par la détermination supplémentaire des coefficients appropriés des sommes complètes.

En général l'évaluation des sommes imbriquées de population finie est le résultat de l'application répétée de la règle

$$\sum_{j_1\neq ...\neq j_t=1}^{N} \left(\prod_{r=1}^{t} x_{i_r j_r}\right) = \sum_{j_1\neq ...\neq j_{t-1}=1}^{N} \left[\prod_{r=1}^{t-1} x_{i_r j_r} \left(\sum_{j_t=1}^{N} x_{i_t j_t}\right)\right] - \sum_{j_1\neq ...\neq j_{t-1}=1}^{N} \left[\sum_{l=1}^{t-1} x_{i_t j_l} \left(\prod_{r=1}^{t-1} x_{i_r j_r}\right)\right]. \quad (21)$$

Cette expression imite la règle d'inclusion-exclusion selon laquelle le premier ensemble de sommes à la droite suit la partie exclusion de la règle et le deuxième ensemble suit la

forme de somme de produits, en particulier $\sum_{j\neq k=1}^{N} x_{i_1 j} x_{i_1 k} = \sum_{j=1}^{N} x_{i_1 j} \sum_{j=1}^{N} x_{i_1 j} - \sum_{j=1}^{N} x_{i_1 j} x_{i_1 j}$, la troisième opération donne

$$E[\{(m(x_{i_1})\}^2] = \frac{N(n-1)}{(N-1)n}\{M(x_{i_1})\}^2 + \frac{N-n}{n(N-1)}M(x_{i_1}^2). \quad (13)$$

Dans (13) $M(x_{i_1}) = K(x_{i_1})$ et $M(x_{i_1}^2) = [N/(N-1)]K(x_{i_1}, x_{i_1}) + K(x_{i_1})K(x_{i_1})$ de sorte que (13) peut s'exprimer de nouveau sous la forme

$$E(m(x_{i_1})^2) = \{K(x_{i_1})\}^2 + (N-n)K(x_{i_1}, x_{i_1})/(Nn). \quad (14)$$

De même, si l'on suit le schéma qui se trouve dans (10) les opérations qui permettent de trouver un estimateur non biaisé de, par exemple, $\{M(x_{i_1})\}^2$ ressemblent à (11), (12) et (13). L'estimande $\{M(x_{i_1})\}^2$ est exprimé dans des sommes imbriquées semblables à (11). Ces sommes seront des sommes imbriquées de population finie. Comme pour (12) on applique les probabilités d'inclusion. Dans ce cas, les sommes de population finie sont remplacées par des sommes d'échantillons et l'opérande est divisée par la probabilité d'inclusion appropriée. Enfin, comme pour (13) les sommes imbriquées de l'échantillon qui en résultent s'expriment sous forme de produits de sommes.

Chacune des opérations élémentaires qui servent à obtenir une valeur attendue par l'entremise des équations (11), (13) et (14), ou à obtenir un estimateur non biaisé, peut être effectuée à l'aide de partitions. Ces opérations sont: l'expression de sommes de produits sous forme de sommes imbriquées et inversement et l'expression de moyennes sous formes de statistiques k et inversement.

4. PARTITIONS ET PROCÉDURES FONDAMENTALES

Un aspect central de l'automatisation de tous les calculs algébriques considérés ici est la notion de partition. Le partitionnement comme point de convergence donne l'impression que les méthodes automatisées qui sont présentées ici ne sont rien de plus que la partition de nombres entiers ou la partition d'un ensemble d'indices. Nous pouvons supposer que la partition d'un nombre entier est comprise, mais une partition complète suppose une définition plus poussée.

Soit $I_m = \{i_1, ..., i_m\}$ un ensemble d'indices m. Une partition simple P_m de I_m divise les indices m en $k \le m$ sous-ensembles ou blocs mutuellement exclusifs et exhaustifs de I_m. Nous écrivons $P_m = (b_1 \mid b_2 \mid ... \mid b_k)$, où les $b_1, ..., b_k$ sont les blocs de I_m. P_m est unique jusqu'aux permutations des indices à l'intérieur des blocs b_i. Le bloc b_i est constitué d'un sous-ensemble des indices de I_m. Des éléments à l'intérieur d'un bloc peuvent se limiter à un ordre alphabétique et les blocs eux-mêmes peuvent être ordonnés de façon que les premiers éléments de chaque bloc soient dans l'ordre alphabétique. On assure ainsi le caractère unique de la partition P_m. Dans ce cas P_m serait une partition ordonnée standard. Le fait d'ordonner les partitions de cette façon n'offre aucun avantage du point de vue des calculs et ne représente donc pas une exigence pour ce qui suit. La partition complète de I_m est l'ensemble $\wp_m$ de toutes les partitions simples P_m de I_m.

Nous pouvons maintenant cerner la partition complète de I_m par voie d'algorithme à l'aide d'une règle d'inclusion-exclusion.

i. Soit $\wp_1 = \{i_1\}$.
ii. Une règle d'inclusion-exclusion détermine la contribution à $\wp_t$ par une partition $P_{t-1} \in \wp_{t-1}$. Dans la partie inclusion de la règle, le nouvel indice i_t est ajouté tour à tour comme un élément à chacun des blocs $b_1, ..., b_k$ qui constituent P_{t-1}. Si P_{t-1} comporte k blocs, on crée k partitions pour $\wp_t$. Dans la partie exclusion de la règle, un nouveau bloc contenant l'indice simple i_t est ajouté à P_{t-1}.

Par exemple, la partition complète de $I_3 = \{i_1, i_2, i_3\}$ est fournie par les étapes

i. $\wp_1 = \{(i_1)\}$
ii. $\wp_2 = \{(i_1 i_2), (i_1 \mid i_2)\}$ (15)
iii. $\wp_3 = \{(i_1 i_2 i_3), (i_1 i_2 \mid i_3), (i_1 i_3 \mid i_2), (i_1 \mid i_2 i_3), (i_1 \mid i_2 \mid i_3)\}$.

De l'étape (i) à l'étape (ii) la règle d'inclusion donne la partition $(i_1 i_2)$ et la règle d'exclusion donne $(i_1 \mid i_2)$. De l'étape (ii) à l'étape (iii) la règle d'inclusion entraîne la création des partitions $(i_1 i_2 i_3)$, $(i_1 i_3 \mid i_2)$, et $(i_1 \mid i_2 i_3)$. La règle d'exclusion donne les partitions $(i_1 i_2 \mid i_3)$ et $(i_1 \mid i_2 \mid i_3)$. Ce type de construction est facile à automatiser puisqu'il dépend d'une règle simple. On trouvera dans Stafford (1996) des renseignements détaillés sur l'automatisation de la partition des indices en partitions complètes et en partitions d'ensembles complémentaires.

Considérons, par exemple, le problème classique de l'écriture des moments suivant un modèle du vecteur aléatoire x_{i_1} en fonction de ses cumulants. Comme dans (5) nous pouvons identifier le h-ième tableau des moments en différenciant FGM($\boldsymbol{t}$) dans (4) h fois et en posant $\boldsymbol{t}$ égal au vecteur zéro. Le résultat est le h-ième coefficient du développement de FGM($\boldsymbol{t}$). Par équivalence, on peut appliquer la même opération à $\exp\{\mathrm{K}(\boldsymbol{t})\}$. Dans ce cas, le résultat est une somme qui dépend des coefficients de $\mathrm{K}(\boldsymbol{t})$ dans (6). Par exemple, on peut écrire les trois premiers moments sous forme de cumulants comme suit:

$$\begin{aligned}
\mu_{i_1} &= \kappa_{i_1} \\
\mu_{i_1 i_2} &= \kappa_{i_1 i_2} + \kappa_{i_1}\kappa_{i_2} \\
\mu_{i_1 i_2 i_3} &= \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2}\kappa_{i_3} + \kappa_{i_1 i_3}\kappa_{i_2} + \kappa_{i_1}\kappa_{i_2 i_3} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3}.
\end{aligned}$$

Dans chaque cas donc le résultat est une somme pour les partitions complètes indiquées dans (15). Ces partitions surviennent car la règle de multiplication pour la différenciation imite la règle d'inclusion-exclusion pour l'énumération de la partition complète.

Le résultat ci-dessus est appliqué à la théorie de l'échantillonnage où nous examinons le problème de la détermination de la valeur espérée d'un produit de sommes d'échantillons. Le calcul nécessite le développement du

est la fonction qui produit des cumulants, où

$$\kappa_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \ldots \partial t_{i_h}} \mathrm{K}(\boldsymbol{t})\big|_{t=0}.$$

Les statistiques k de population finie, notées $K(\cdot)$, se définissent comme les estimateurs non biaisés (dans le cadre du modèle de superpopulation i.d.i.) des cumulants de modèle associés. Le nombre d'arguments dans K séparés par des virgules indique l'ordre de la statistique k. Ainsi, la statistique k d'ordre trois $K(x_{i_1}, x_{i_2}, x_{i_3})$ est l'estimation non biaisée basée sur le modèle de (6), où

$$K(x_{i_1},x_{i_2},x_{i_3}) = \frac{N}{(N-1)(N-2)} \times \sum_{j\in U} [x_{i_1 j} - M(x_{i_1})][x_{i_2 j} - M(x_{i_2})][x_{i_3 j} - M(x_{i_3})]. \quad (7)$$

Dans le cas à une variable les statistiques k de population finie sont décrites dans Wishart (1952). En particulier $K(y, y)$ et $K(y, y, y)$ dans la notation courante sont K_2 et K_3 dans la notation de Wishart (1952). Les statistiques k de l'échantillon, notées $k(\cdot)$ avec les arguments appropriés, se définissent comme les estimateurs non biaisés dans le cadre d'un sondage aléatoire simple sans remise des statistiques k de population finie associées. Comme dans Wishart (1952) la statistique k de l'échantillon peut être obtenue de la statistique k de la population si l'on remplace N par n et si l'on prend la somme sur les $j \in s$ au lieu de toutes les unités de la population finie. Par exemple,

$$k(x_{i_1}, x_{i_2}, x_{i_3}) = \frac{n}{(n-1)(n-2)} \times \sum_{j\in s} [x_{i_1 j} - m(x_{i_1})][x_{i_2 j} - m(x_{i_2})][x_{i_3 j} - m(x_{i_3})].$$

À noter que s'il n'y a pas de virgule dans la statistique k de la population ou de l'échantillon, le produit des éléments qui apparaissent ensemble est nécessaire. Par exemple, $K(xy)$ est la statistique k de population finie d'ordre un d'une nouvelle variable qui est le produit des mesures x_j et y_j pour $j = 1, ..., N$; $K(x, y)$ est une statistique k d'ordre deux, en particulier la covariance de population finie entre x et y.

3. OPÉRATEURS

L'opérateur d'espérance E peut être appliqué directement à toute somme imbriquée de l'échantillon de façon à fournir une somme imbriquée de population finie. De même, un estimateur non biaisé de toute somme imbriquée de population finie est une somme imbriquée de l'échantillon. Pour ce qui est des sommes imbriquées triples, par exemple,

$$E\Big[\sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l}\Big] = \sum_{J_3 = 1}^{N} \pi_{jkl} x_{i_1 j} x_{i_2 k} x_{i_3 l} \quad (8)$$

et

$$\sum_{J_3 = 1}^{N} x_{i_1 j} x_{i_2 k} x_{i_3 l} \sim \sum_{J_3 \in s} x_{i_1 j} x_{i_2 k} x_{i_3 l} / \pi_{jkl}, \quad (9)$$

où J_3 est un ensemble d'indices $\{j, k, l\}$ tel que $j \neq k \neq l$ et où π_{jkl} est une probabilité d'inclusion conjointe. Des expressions parallèles peuvent être établies pour des schémas d'échantillonnage avec remise.

À noter que m sera non biaisé pour le M associé dans le cadre d'un sondage aléatoire simple sans remise. En général, pour tout plan d'échantillonnage de taille fixe n,

$$E[m(x_{i_1} x_{i_2} x_{i_3})] = \frac{N}{n} M(x_{i_1} x_{i_2} x_{i_3} \pi)$$

et

$$M(x_{i_1} x_{i_2} x_{i_3}) \sim \frac{n}{N} m(x_{i_1} x_{i_2} x_{i_3} / \pi)$$

où $M(x_{i_1} x_{i_2} x_{i_3})$ et $m(x_{i_1} x_{i_2} x_{i_3})$ sont définis dans (1) et (2) respectivement.

La détermination de l'espérance d'un estimateur $\hat{\theta}$ ou la détermination d'un estimateur non biaisé pour le paramètre de θ peut être représentée schématiquement sous la forme

$$\sum\prod \Rightarrow \sum\sum \Rightarrow \sum\prod, \quad (10)$$

où $\sum\prod$ signifie la somme des produits et $\sum\sum$ signifie une somme de sommes imbriquées. Si θ ou $\hat{\theta}$ peut s'exprimer comme une quantité $\sum\prod$, c.-à-d. une somme de produits de moyennes, la détermination d'un estimateur non biaisé de θ ou de moments de $\hat{\theta}$ se réduit à suivre le schéma que l'on trouve dans (10) et à appliquer l'opérateur approprié, comme ceux qui sont donnés dans (8) ou (9), à $\sum\sum$, l'étape moyenne du schéma. Si θ ou $\hat{\theta}$ sont des fonctions continues de moyennes, mais ne se laissent pas exprimer directement comme des quantités $\sum\prod$, il faut une étape initiale avant de pouvoir appliquer le schéma dans (10). Pour $\hat{\theta}$ l'étape initiale consiste à obtenir un développement en série de Taylor. Pour θ l'étape initiale consiste à obtenir une équation d'estimation puis à résoudre cette équation pour le paramètre.

Illustrons le schéma que l'on trouve dans (10) en considérant le cas simple de la détermination de $E[\{m(x_{i_1})\}^2]$ dans le cadre d'un sondage aléatoire simple sans remise. Il s'agit d'abord d'exprimer $\{m(x_{i_1})\}^2$ sous forme de sommes imbriquées. En particulier,

$$\{m(x_{i_1})\}^2 = \frac{1}{n^2} \sum_{j\in s} x_{i_1 j}^2 + \frac{1}{n^2} \sum_{j\neq k \in s} x_{i_1 j} x_{i_1 k}. \quad (11)$$

Nous avons là l'étape $\sum\prod \Rightarrow \sum\sum$. L'opérateur d'espérance peut maintenant être appliqué à $\sum\sum$. Lorsqu'on applique les probabilités d'inclusion $\pi_j = n/N$ et $\pi_{jk} = n(n-1)/[N(N-1)]$, l'opération d'espérance sur (11) donne

$$\frac{1}{n^2} \frac{n}{N} \sum_{j=1}^{N} x_{i_1 j}^2 + \frac{1}{n^2} \frac{n(n-1)}{N(N-1)} \sum_{j\neq k=1}^{N} x_{i_1 j} x_{i_1 k}. \quad (12)$$

L'étape $\sum\sum \Rightarrow \sum\prod$ est maintenant appliquée. Lorsqu'on exprime la somme imbriquée qui se trouve dans (12) sous

en résulte la possibilité d'automatiser le calcul. Des formules apparemment sans lien peuvent résulter de la même règle fondamentale et un même outil d'algèbre informatique peut servir à mettre en oeuvre de nombreux calculs différents.

La notation utilisée dans le présent exposé est expliquée à la section 2. On trouvera à la section 3 une analyse des opérateurs d'espérance. La concept de partitionnement est passé en revue à la section 4 et on y trouve une règle qui donne lieu à une méthode récursive simple d'énumération des partitions. Il est question à la section 5 des partitions de nombres entiers et de la linéarisation de Taylor. La section 6 indique comment l'énumération des partitions entraîne le calcul automatique des valeurs attendues de produits de moyennes d'échantillon et de statistiques k de même que le développement d'estimateurs non biaisés de produits de moyennes de populations finies et de statistiques k. Dans cette même section, nous appliquons la méthode à l'estimation de quotients et de la régression.

L'automatisation de ces calculs et développements donne lieu à des procédures que l'on peut exécuter instantanément et sans erreur à l'aide d'un ordinateur. De même, le recours à des formules possiblement longues et complexes est éliminé. Bon nombre de calculs algébriques manuels peuvent ainsi être évités. Les codes machine de mise en oeuvre de la méthode décrite ici ont été rédigés en langage symbolique *Mathematica* 2.0 monté sur un IBM Risc 6000 muni de 64 mégaoctets de mémoire vive. Celui-ci est accessible par FTP anonyme au *fisher.stats.uwo.ca*. Même si nous utilisons *Mathematica*, il est sans doute possible de procéder dans d'autres environnements comme *Maple*, *Macsyma* ou *Reduce*. Ainsi, Kendall (1993) décrit un système, mis en oeuvre dans *Reduce*, qui permet d'identifier des expressions invariantes. On trouvera dans Kendall (1993) un aperçu complet de l'algèbre informatique en probabilité et en statistique avant 1991.

2. ÉLÉMENTS DE NOTATION

Soit une population finie de taille N. Une mesure d'intérêt y_j est prise sur chaque unité $j, j \in U = \{1, ..., N\}$. De plus, une variable auxiliaire simple x_j ou possiblement un vecteur $P \times 1$ de variables auxiliaires $\boldsymbol{x}_j$ peut être pris sur les unités. La p-ième entrée de ce vecteur $\boldsymbol{x}_j$ est x_{pj}, où $p = 1, ..., P$. Plusieurs types de paramètres de population finie peuvent être définis sur les mesures y_j, x_j, ou $\boldsymbol{x}_j$ pour $j = 1, ..., N$. Nous notons θ un paramètre d'intérêt de population finie. Il est souvent possible d'exprimer θ sous forme de fonction continue de moyennes de population finie, de moments centrés et de statistiques k. Par souci de simplicité ici nous traitons uniquement de moyennes et de statistiques k. À noter que les variances et covariances de population finie sont également des statistiques k d'ordre deux.

Les éléments de population N ne sont pas tous observés. Supposons qu'un échantillon s de taille n soit choisi dans une population U selon un schéma d'échantillonnage quelconque. Un estimateur de θ, donné par $\hat{\theta}$, est une fonction continue de moyennes d'échantillons et de statistiques k d'échantillons.

Afin d'éviter une bonne partie de la notation encombrante de sommation, nous adaptons la notation des indices de McCullagh (1987) à nos besoins. Pour tout j le vecteur $\boldsymbol{x}_j$ contient des entrées P de façon que chacune de ces variables x puisse être associée à l'un des indices P. Soit $\{i_1, ..., i_m\}$ un sous-ensemble de m de ces indices P. Dans notre adaptation de la notation de McCullagh, $x_{i_1 j}$ représente maintenant ce que nous avons appelé le vecteur $\boldsymbol{x}_j$. Les produits de ces quantités en indice deviennent des tableaux à plusieurs dimensions. Ainsi, le produit $x_{i_1 j} x_{i_2 j} x_{i_3 j}$ est un tableau tridimensionnel de dimension $P \times P \times P$.

Notons M une moyenne de population finie. L'argument de M indique la structure de l'opérande dans la moyenne. Par exemple, $M(y) = \sum_{j \in U} y_j / N$ et $M(yy)$ ou par équivalence $M(y^2) = \sum_{j \in U} y_j^2 / N$. Dans la notation des indices, par exemple,

$$M(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in U} x_{i_1 j} x_{i_2 j} x_{i_3 j} / N \tag{1}$$

est un tableau tridimensionnel. Un élément de ce tableau est la moyenne de produits dans l'une des permutations des éléments P pris trois à la fois dans $\boldsymbol{x}_j$ où jusqu'à trois des éléments peuvent être semblables. Le (p, q, r)-ième élément de ce tableau est $\sum_{j \in U} x_{pj} x_{qj} x_{rj}$ où $p, q, r = 1, ..., P$. La moyenne de l'échantillon est notée m de sorte que, par exemple,

$$m(x_{i_1} x_{i_2} x_{i_3}) = \sum_{j \in s} x_{i_1 j} x_{i_2 j} x_{i_3 j} / n. \tag{2}$$

Afin de réaliser des expansions asymptotiques, puisque la variance d'un estimateur donné $\hat{\theta}$ sera $O(n^{-1})$, nous définissons une variable normalisée pour $\hat{\theta}$: il s'agit de la variable originale $\hat{\theta}$ centrée par rapport à son espérance et réduite par $1/\sqrt{n}$. C'est-à-dire,

$$z(\hat{\theta}) = [\hat{\theta} - E(\hat{\theta})] \sqrt{n}. \tag{3}$$

Au besoin nous utilisons la convention de sommation de McCullagh (1987), où les indices inférieurs répétés comme indices supérieurs indiquent des sommes implicites pour cet indice. Si, par exemple, on suppose que les $\boldsymbol{x}_j$ sont des vecteurs indépendants et distribués de façon identique (i.d.i.) provenant d'une superpopulation infinie, on peut obtenir des moments de superpopulation à plusieurs variables par l'entremise de la fonction génératrice des moments et qui selon cette convention s'exprime comme suit:

$$\mathrm{FGM}(\boldsymbol{t}) = 1 + \sum_{h=1}^{\infty} \mu_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h!, \tag{4}$$

où

$$\mu_{i_1 \cdots i_h} = \frac{\partial^h}{\partial t_{i_1} \ldots \partial t_{i_h}} \mathrm{FGM}(\boldsymbol{t})|_{t=0}. \tag{5}$$

Par définition, la relation entre la fonction génératrice des moments et la fonction génératrice des cumulants est déterminée par la règle $\mathrm{FGM}(\boldsymbol{t}) = \exp\{K(\boldsymbol{t})\}$, où

$$\mathrm{K}(\boldsymbol{t}) = \sum_{h=1}^{\infty} \kappa_{i_1 \cdots i_h} \prod_{j=1}^{h} t^{i_j} / h! \tag{6}$$

Techniques d'enquête, juin 1997
Vol. 23, n° 1, pp. 3-11
Statistique Canada

Une algèbre informatique pour la théorie des enquêtes par échantillonnage

J.E. STAFFORD et D.R. BELLHOUSE[1]

RÉSUMÉ

Les auteurs présentent un système de procédures qui peut servir à automatiser les calculs algébriques complexes que l'on retrouve souvent en théorie des enquêtes par échantillonnage. Ils montrent que trois techniques de base en théorie de l'échantillonnage dépendent de l'application répétée de règles donnant lieu à des partitions: le calcul des valeurs espérées dans un plan d'échantillonnage quelconque à un degré, la détermination d'estimateurs non biaisés ou convergents dans le cadre de ces plans et le développement en séries de Taylor. La méthode est appliquée ici à des calculs de moments de la moyenne de l'échantillon, de l'estimateur de quotients et de l'estimateur de la régression dans le cas spécial d'un sondage aléatoire simple sans remise. L'innovation présentée ici est que les calculs peuvent désormais être exécutés instantanément par ordinateur sans erreur et sans recours à des formules existantes possiblement longues et complexes. Un autre avantage immédiat est que les calculs peuvent être exécutés là où il n'existe actuellement aucune formule. Le code machine élaboré en vue de la mise en oeuvre de cette méthode est accessible par FTP anonyme au *fisher.stats.uwo.ca.*.

MOTS CLÉS: Statistiques *k*; partitions; moments de produits; estimateurs de quotient et de régression; calculs symboliques; estimation de la variance.

1. INTRODUCTION

En théorie classique de l'échantillonnage, deux problèmes généraux nous préoccupent. Il s'agit de la détermination d'un estimateur non biaisé d'un paramètre θ et du calcul des moments de $\hat{\theta}$, l'estimateur de θ.

La méthode de traitement de base des espérances et de l'estimation non biaisée consiste à effectuer, sur l'échantillon et la population, des sommes imbriquées respectivement par l'entremise des probabilités d'inclusion, c'est-à-dire les probabilités simples ou conjointes selon le cas. Une somme imbriquée est une somme couvrant l'étendue d'un ou plusieurs indices de sorte que chaque terme de la somme dépend d'indices de valeur différente. Un estimateur non biaisé d'une somme imbriquée quelconque de population est la somme imbriquée de l'échantillon associé, la quantité sous la sommation étant divisée par la probabilité d'inclusion appropriée. De même, l'espérance d'une somme imbriquée quelconque d'un échantillon est la somme imbriquée de la population associée, la quantité sous la sommation étant multipliée par la probabilité d'inclusion appropriée.

En théorie de l'échantillonnage, comme dans plusieurs autres domaines de la statistique, de nombreux calculs algébriques dépendent d'une partition quelconque. Pour ce qui est de l'échantillonnage en particulier, Wishart (1952) a montré que les calculs de moments de base dans le cadre d'un sondage aléatoire simple sans remise dépendent largement de partitions. Nous utiliserons ici des partitions pour exprimer la somme de produits de moyennes ou de totaux sous forme de combinaisons linéaires de sommes imbriquées et inversement.

Dans les résultats que nous présentons ici, nous considérons la situation dans laquelle θ et $\hat{\theta}$ peuvent être exprimés sous forme de fonctions continues de moyennes ou de totaux (population ou échantillon selon le cas). Il existe deux possibilités: la fonction continue en question peut être exprimée comme la somme de produits de moyennes ou de totaux, ou bien la fonction continue ne peut pas être exprimée de cette façon. Lorsque la deuxième possibilité s'applique, la fonction $\hat{\theta}$ est d'abord linéarisée par un développement en série de Taylor et θ est exprimé sous forme de racine d'une équation d'estimation. Nous utilisons des partitions de nombres entiers afin d'obtenir les termes de la linéarisation de Taylor d'une fonction ou pour la racine d'une fonction. Le résultat final est que θ et $\hat{\theta}$ peuvent être exprimés, exactement ou approximativement, comme la somme de produits de moyennes ou de totaux qui, à leur tour, peuvent être exprimés sous forme de combinaisons linéaires de sommes imbriquées et inversement. L'estimation de θ ou le calcul des moments de $\hat{\theta}$ représente dès lors une procédure en trois étapes: a) exprimer une équation d'estimation pour θ ou l'estimateur $\hat{\theta}$ comme la somme de produits de moyennes ou de totaux, en utilisant au besoin la linéarisation de Taylor; b) transformer l'expression obtenue à la première étape en une combinaison linéaire de sommes imbriquées et traiter ces sommes imbriquées de façon à obtenir des estimations ou des espérances non biaisées selon le cas; c) transformer les sommes imbriquées résultant de la deuxième étape de nouveau en une somme de produits de moyennes ou de totaux.

La clé de l'automatisation des résultats de la théorie de l'échantillonnage est d'avoir recours à des partitions. En général, les partitions simples comme celles d'un nombre entier aussi bien que les partitions plus complexes comme les partitions complètes sont le résultat de l'application répétée d'une règle fondamentale. Lorsque la règle est identifiée, il

[1] J.E. Stafford et D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London (Ontario), N6A 5B7.

Dans son article, Losinger propose un estimateur modifié de l'erreur-type des groupes aléatoires pour les données obtenues de l'échantillon du recensement décennal aux États-Unis. L'estimateur habituel des groupes aléatoires comporte deux propriétés non souhaitables pour les variables binomiales: premièrement, les estimations de l'erreur-type pour les réponses «oui» et «non» ne sont pas égales; deuxièmement, si tous les répondants indiquent «oui», l'erreur-type estimée n'est pas égale à zéro. La modification proposée consiste essentiellement à appliquer un ajustement de ratio à chaque estimation par sous-groupe afin qu'il y ait concordance entre les estimations par sous-groupe de la population et la valeur totale.

Enfin, Zeelenberg présente une technique simple qui fait appel à l'utilisation des différentiels pour linéariser des estimateurs non linéaires basés sur le plan. En bout de ligne, les expressions linéarisées permettent d'obtenir des expressions simples basées sur la méthode de Taylor pour les variances des estimateurs non linéaires. L'auteur illustre la méthode proposée à l'aide de deux exemples : l'estimateur du coefficient de régression et l'estimateur de régression.

Le rédacteur en chef

Techniques d'enquête, juin 1997
Vol. 23, N° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles qui traitent de divers sujets. Dans le premier article, Stafford et Bellhouse présentent les blocs fonctionnels de base pour l'élaboration d'une algèbre informatique complète pour la théorie de l'échantillonnage. Ils montrent que trois techniques de base de la théorie de l'échantillonnage dépendent de l'application répétée des règles qui donnent lieu aux partitions. La méthodologie est illustrée au moyen d'applications au calcul des moments de la moyenne de l'échantillon, de l'estimateur par quotient et de l'estimateur de régression dans le cas spécial de l'échantillonnage aléatoire simple sans remise. L'application machine de la méthodologie décrite a été faite à l'aide du langage de programmation *Mathematica*.

Hinkins, Oh et Scheuren exposent une nouvelle stratégie pour l'analyse des données d'enquêtes complexes. Ils ont prélevé un sous-échantillon de telle manière que le sous-échantillon puisse être considéré comme un échantillon aléatoire simple de la population initiale, puis ils y ont appliqué la méthode standard pour les variables aléatoires indépendantes de même distribution. Ils proposent de répéter la méthode plusieurs fois pour récupérer l'information perdue durant le sous-échantillonnage de l'échantillon initial. Ils montrent comment appliquer leur méthode à l'échantillonnage stratifié, à l'échantillonnage en grappes à un ou deux degrés ainsi qu'aux plans avec deux unités primaires d'échantillonnage par strate.

Nascimento Silva et Skinner examinent le problème de la sélection des variables pour l'estimation de régression. Ces auteurs ont mis au point une méthode axée sur la réduction de l'erreur quadratique moyenne de l'estimateur obtenu. Ils comparent de façon empirique leur approche à d'autres en utilisant les données d'un essai fait en 1988 sur les techniques brésiliennes de recensement; les méthodes proposées ont et de bonnes propriétés en ce qui a trait à l'erreur quadratique moyenne et au biais.

Eltinge et Yansaneh étudient le problème de la formation de cellules d'ajustement pour la non-réponse. Dans le contexte des paradigmes généraux des cellules basées sur l'estimation des unités et des probabilités, ils examinent une variété de diagnostics pour l'évaluation d'un ensemble de cellules d'ajustement. Les méthodes de diagnostic examinées incluent les suivantes : comparaison des estimations et des erreurs-types pour différents nombres de cellules d'ajustement; évaluation du biais à l'intérieur des cellules; évaluation de l'étendue des cellules en regard de la précision des probabilités de réponse estimées et comparaisons des estimations basées sur les cellules par rapport aux estimations non corrigées.

Kovačević et Yung ont mené une étude empirique visant à comparer des méthodes d'estimation de la variance pour les mesures de l'inégalité du revenu estimées à partir de données d'enquêtes complexes. Les méthodes d'estimation de la variance examinées sont les suivantes: méthode jackknife, méthode «bootstrap», méthode du demi-échantillon équilibré groupé, méthode itérative du demi-échantillon équilibré groupé et méthode de Taylor basée sur les équations d'estimation. Après avoir comparé les biais relatifs, la stabilité relative et les propriétés de couverture des intervalles de confiance associés, pour un certain nombre de mesures de l'inégalité du revenu, ils concluent que la méthode de Taylor est celle qui donne les meilleurs résultats, suivie de la méthode «bootstrap».

Humphreys et Skinner étudient l'utilisation de la méthode de la variable instrumentale pour l'estimation des mouvements bruts entre états discrets. Cette méthode pourrait être utile lorsque les estimations externes des taux de classification erronée ne sont pas disponibles. Ils illustrent leur méthode à l'aide de données obtenues de la U.S. Panel Study of Income Dynamics, les deux états utilisés étant «actifs» et «inactifs». Leurs résultats indiquent que, lorsqu'une erreur de mesure est présente, les estimations non corrigées peuvent comporter un biais considérable – un problème que l'on peut éviter en utilisant des variables instrumentales appropriées.

Waksberg, Judkins et Massey discutent des problèmes liés au suréchantillonnage de régions géographiques pour produire des estimations pour de petits domaines de la population lors d'enquêtes démographiques, conjointement avec la présélection des ménages. Ils présentent une évaluation empirique de la réduction de la variance, ainsi qu'une évaluation de la robustesse de l'échantillonnage dans le temps. Ils discutent du suréchantillonnage géographique simultané pour l'estimation de plusieurs petits domaines.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 23, numéro 1, juin 1997

TABLE DES MATIÈRES

Dans ce numéro 1

J.E. STAFFORD et D.R. BELLHOUSE
Une algèbre informatique pour la théorie des enquêtes par échantillonnage 3

S. HINKINS, H.L.OH et F. SCHEUREN
Algorithmes de plan de sondage inverses 13

P.L.D. NASCIMENTO SILVA et C.J. SKINNER
Sélection des variables pour l'estimation par régression dans le cas des populations finies 25

J.L. ELTINGE et I.S. YANSANEH
Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey 37

M.S. KOVAČEVIĆ et W. YUNG
Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique 47

K. HUMPHREYS et C.J. SKINNER
Estimation des variables instrumentales des flux bruts en présence de l'erreur de mesure 61

J. WAKSBERG, D. JUDKINS et J.T. MASSEY
Suréchantillonnage géographique dans les enquêtes démographiques aux États-Unis 69

W.C. LOSINGER
Nouvel estimateur de l'erreur-type pour les groupes aléatoires 81

K. ZEELENBERG
Dérivation simple de l'estimateur de régression par linéarisation 85

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 $ par année au Canada et de 47 $ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1997 • VOLUME 23 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

Juillet 1997

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Nº 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 1997

•

VOLUME 23

•

NUMÉRO 1

Statistique Canada
Statistics Canada

Canada

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1997

•

VOLUME 23

•

NUMBER 2

Statistics Canada   Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1997 • VOLUME 23 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

February 1998

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 23, Number 2, December 1997

CONTENTS

In This Issue 79

P.S. KOTT and D.M. STUKEL
Can the Jackknife Be Used With a Two-Phase Sample? 81

G. DECAUDIN and J.-C. LABAT
A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France 91

P. RAVALET
An Adaptive Procedure for the Robust Estimation of the Rate of Change of Investment 99

F. COTTON and C. HESSE
Sampling and Maintenance of a Stratified Panel of Fixed Size 109

P.J. FARRELL
Empirical Bayes Estimation of Small Area Proportions Based on Ordinal Outcome Variables 119

A. GELMAN and T.C. LITTLE
Poststratification Into Many Categories Using Hierarchical Logistic Regression 127

K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN and G. NARAYANA
Estimating the Population and Characteristics of Health Facilities and Client Populations Using a Linked Multi-Stage Sample Survey Design 137

J. DUFOUR, R. KAUSHAL and S. MICHAUD
Computer-assisted Interviewing in a Decentralised Environment: The Case of Household Surveys at Statistics Canada 147

F. SCHEUREN and W.E. WINKLER
Regression Analysis of Data Files That Are Computer Matched - Part II 157

Acknowledgements 167

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 79-80
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Kott and Stukel consider jackknife variance estimation for a specific, but widely used two-phase design. At the first phase, clusters within strata are selected using SRS with replacement, and all units within the selected clusters are sampled. At the second phase, the sampled units are restratified and then second phase units are selected using SRS without replacement. Two point estimators are considered: the "reweighted expansion estimator" and the more commonly known "double expansion estimator". Under this design, it is shown that the jackknife variance estimator behaves remarkably better for the former point estimator than it does for the latter. A Monte Carlo study supports these findings.

Decaudin and Labat describe a "multi-source" population estimation system designed to produce local population estimates during intercensal periods in France. The system is robust and flexible in that it works with a variable number of sources. It is based on a robust combination of estimates from different sources, blending demographic reasoning with statistical methods.

Ravalet applies GM-estimators to INSEE's industrial investment survey with an adaptive procedure to produce a robust estimator. Tukey's biweight function and the Cauchy function are examined. Each function relies on a tuning constant based on the width of the tail of the distribution and the concentration of the residuals. Tuning constants that minimize the estimator's variance are determined for eight distributions representing various scenarios relating to the width of the tail and the concentration of the residuals, which are assumed to be symmetrical.

Cotton and Hesse study the characteristics of various methods of selecting a stratified panel of fixed size, along with their impact on initial selection, rotation, resampling and sample overlap. The authors propose a kind of algorithm based on transformations of permanent random numbers used for sampling purposes; the algorithm extends the pre-resampling rotation into the post-resampling period. The transformations can be performed on random numbers that have been made equidistant and on random numbers derived from a uniform distribution.

In his paper Farrell studies empirical Bayes estimation of small area proportions. Using data from the United States Census he compares empirical Bayes small area estimates of proportions of individuals in different income categories based on multinomial and ordinal logistic models with random effects. Inferences based on the ordinal model were slightly better than those based on the multinomial model. He also compares naive and bootstrap adjusted variance estimates and coverage probabilities of their associated confidence intervals. The bootstrap adjustment improves coverage significantly.

Gelman and Little describe a novel extension of analyzing poststratified survey data, using Bayesian hierarchical logistic regression modelling. The technique allows for many more stratification categories than are typically feasible using standard poststratification and weighting strategies, and thus much more population level information can be included in the model. The proposed method as well as some of the more standard methods are applied to pre-election opinion polling data in the U.S., and the various models are evaluated graphically by comparing them to actual election outcomes.

Singh, Tsui, Suchindran and Narayana describe the survey design and estimation techniques used for PERFORM (Project Evaluation Review for Organizational Resource Management), a large scale survey conducted in the state of Uttar Pradesh in India. The survey was designed to estimate the characteristics of health facilities and their target populations, in order to provide benchmark indicators for a large family planning project. PERFORM uses a stratified multi-stage design, where the ultimate sampling units are households and eligible females residing within. However, estimates of health facilities, which are not explicitly part of the sampling scheme, are also obtained by adjusting for multiplicity of the selected secondary sampling units served by those health facilities.

Dufour, Kaushal and Michaud review the tests and studies that preceeded the implementation of computer-assisted interviewing for most household surveys at Statistics Canada. The interviewing is conducted, in person at the respondent's home or by telephone from the interviewer's home, using laptop computers. They also discuss the challenges that were faced with the implementation of the new technology into ongoing surveys and the new opportunities for monitoring survey collection offered by it.

Scheuren and Winkler propose a method for using noncommon but correlated quantitative variables to improve record linkage. The basic idea is to use the linkages which are almost certainly correct to estimate a regression relationship between the noncommon variables and then to use the predicted values of these variables in a subsequent record linkage step. The procedure can then be iterated until convergence. The regression step uses a procedure which adjusts the regression for possible errors in the linkage, described in an article by the same authors in the June 1993 issue of *Survey Methodology*. The method is illustrated empirically and it is shown that it can lead to good results in situations that were hitherto hopeless.

The Editor

Dear *Survey Methodology* Reader,

I would like to take a moment to thank you for your interest and support of *Survey Methodology*. Since its inception, the journal remains committed to publishing articles relevant to statistical agencies and researchers with emphasis on the development and evaluation of specific methodologies as applied to data collection or to the data themselves.

Survey Methodology is approaching its 25th anniversary. From its beginning as an in-house review of developments in survey methodology in Statistics Canada, it has evolved into a widely read statistical journal with an editorial board of internationally recognized survey statisticians. Though many improvements to content and presentation have occurred during this period, there is always room for improvement. I would appreciate any suggestions, comments and recommendations you may have to assist us in our task of maintaining *Survey Methodology* as a viable platform for statistical development into the next millennium.

Should you wish to have complimentary copies of *Survey Methodology* sent to a colleague, please do not hesitate to contact us.

I thank you again for your interest and continued support of *Survey Methodology*.

Sincerely,

M.P. Singh
singhmp@statcan.ca

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 81-89
Statistics Canada

Can the Jackknife Be Used With a Two-Phase Sample?

PHILLIP S. KOTT and DIANA M. STUKEL[1]

ABSTRACT

The jackknife variance estimator has been shown to have desirable properties when used with smooth estimators based on stratified multi-stage samples. This paper focuses on the use of the jackknife given a particular two-phase sampling design: a stratified with-replacement probability cluster sample is drawn, elements from sampled clusters are then restratified, and simple random subsamples are selected within each second-phase stratum. It turns out that the jackknife can behave reasonably well as an estimator for the variance for one common "expansion" estimator but not for another. Extensions to more complex estimation strategies are then discussed. A Monte Carlo study supports our principal findings.

KEY WORDS: Stratified; Reweighted expansion estimator; Double expansion estimator; Asymptotic.

1. INTRODUCTION

Krewski and Rao (1981) and Rao and Wu (1985) explore the design-based properties of the jackknife variance estimator given a stratified multi-stage sample incorporating with-replacement sampling in the first stage. Their results, although fairly general, cannot be directly applied to many multi-phase sampling designs. See also Wolter (1985; Chapter 4.5).

In this paper, we consider a simple example of two-phase sampling. A stratified with-replacement probability cluster sample is selected in a first phase of sampling. The elements in sampled clusters are then restratified, perhaps using information gathered from the first-phase sample, and a stratified simple random subsample is drawn without replacement.

One can estimate a total without auxiliary information in one of two ways. In the *double expansion estimator* – called "the π^* estimator" in Särndal, Swensson, and Wretman (1992, p. 347) – the value of each subsampled element is simply multiplied by the product of its expansion factor at each phase (*i.e.*, the inverses of its first-phase and second-phase selection probabilities) and then summed.

Although the double expansion estimator is more easily located in text books, the *reweighted expansion estimator* may be more common in practice, especially when element nonresponse is treated as a second phase of sampling, as in the weighting class estimator of Oh and Scheuren (1983, p. 150). An estimator for the population size of each second-phase stratum is computed by summing the first-phase expansion factors of all the elements in the second-phase stratum before subsampling. This value is then multiplied by the estimated second-phase stratum mean based on the subsample to yield an estimated stratum total. The second-phase estimated stratum totals are finally added together to produce the reweighted expansion estimator for the population total.

We are more concerned here with real two-phase sampling, rather than the artifice of treating nonresponse as an additional sampling phase. The National Agricultural Statistics Service (NASS) presently uses the double expansion estimator in its Quarterly Agricultural Surveys (QAS). A stratified area cluster sample is enumerated in June. Farms identified in the June survey are restratified based on their June responses and then subsampled for enumeration in September, December, and March.

NASS uses a two-phase design and the reweighted expansion estimator for its on-farm chemical use surveys. The first phase of sampling identifies farms with specific crops, and the second phase measures pesticide use on those crops.

This paper shows that although the jackknife may be used to estimate the variance of the reweighted expansion estimator under certain conditions, it is not generally effective as a variance estimator for the double expansion estimator. Section 2 introduces the reweighted expansion estimator and discusses its mean squared error. Section 3 shows that the jackknife variance estimator can be nearly unbiased for the reweighted variance estimator, while Section 4 addresses the jackknife's failings as a variance estimator for the double expansion estimator. Section 5 describes a simulation study that appears to confirm the main assertions of the previous sections. Section 6 discusses extensions of the reweighted expansion estimator, and Section 7 offers some concluding remarks. An appendix provides an outline of our assumed asymptotic framework and some proofs.

2. THE REWEIGHTED EXPANSION ESTIMATOR

2.1 The Estimator

Let $h(=1, ..., H)$ denote the first-phase strata of a stratified with-replacement probability cluster sample, n_h the number of sampled clusters in stratum h, and F_h the set of those clusters. Let $g(=1, ..., G)$ be the second-phase

[1] Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030; Diana M. Stukel, Household Survey Methods Division, Statistics Canada, Ottawa, Canada K1A OT6.

strata from which a stratified simple random subsample is drawn without replacement. An element in a cluster sampled p times in the first phase is treated as p distinct elements for the subsample. Let M_g be the number of elements in g before subsampling and m_g the number of subsampled elements in g. In practice, the G second-phase strata are often not defined until after the first-phase sample has been drawn.

Let S_g be the set of elements in g before subsampling, s_g the set of subsampled elements in g, s the entire set of subsampled elements, and $m = \sum_g m_g$ the subsample size. Finally, let y_i be the value of interest for element i, and w_i the first-phase expansion factor for i (*i.e.*, the inverse of the selection probability for the cluster containing i).

The estimator for the population total, T, one would use if all the elements in the first-phase sample were enumerated can be written as

$$t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i y_i. \tag{1}$$

Let the *reweighted expansion estimator* for T be:

$$t_2 = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} (M_g/m_g) w_i y_i}{\sum_{i \in s_g} (M_g/m_g) w_i} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} \right\}. \tag{2}$$

An alternative expression for t_2 is

$$t_2 = \sum_{g=1}^{G} \sum_{i \in s_g} a_i y_i = \sum_{i \in s} a_i y_i, \tag{3}$$

where

$$a_i = \left[\sum_{k \in S_g} w_k \Big/ \sum_{k \in s_g} w_k \right] w_i \text{ for } i \in s_g$$

is the *adjusted weight* for element i. Equation (3) is what gives the reweighted expansion estimator its name.

2.2 Its Mean Squared Error (Some Theory)

Now t_2 is not, in general, an unbiased estimator of T. Nevertheless, under certain mild conditions specified in the appendix, it is a design consistent estimator for T; that is, $\text{plim}_{m \to \infty} (t_2 - T)/T = 0$ (Isaki and Fuller 1982). For the exposition in the text, it suffices to say that the m_g are assumed to be large.

Observe that

$$E[(t_2 - T)^2] = E[(\{t_1 - T\} + \{t_2 - t_1\})^2]$$
$$\approx \text{Var}_1(t_1) + E_1\{E_2[(t_2 - t_1)^2]\},$$

where the subscripts on Var and E denote the phase of sampling. Since the m_g are assumed to be large, $E_2[t_1(t_2 - t_1)] = t_1 E_2(t_2 - t_1) \approx 0$. Also, $E(t_2 - T) = E_1[E_2(t_2 - T)] \approx 0$, and the mean squared error of t_2 is effectively its (asymptotic) variance.

Since first phase of sampling was conducted with replacement, $\text{Var}_1(t_1)$ can, in principle, be estimated by

$$v_{L1} = \sum_{h=1}^{H} (n_h/[n_h - 1])$$
$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i y_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i y_i \right]^2 / n_h \right), \tag{4}$$

where U_{hj} is the set the elements in sampled cluster j of first-phase stratum h. The subscript L denotes "linearization" for historical reasons although there is nothing to linearize in this context. Note that when there is a second phase of sampling, it will generally not be possible to compute v_{L1} in practice.

Now

$$t_2 - t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i \left\{ \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} - \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i r_i}{\sum_{i \in s_g} w_i},$$

where

$$r_i = y_i - \sum_{k \in S_g} w_k y_k \Big/ \sum_{k \in S_g} w_k \text{ for } i \in S_g.$$

It is crucial for the arguments below to realize that r_i has been defined so that $\sum_{i \in S_g} w_i r_i = 0$ for all g.

Continuing,

$$t_2 - t_1 \approx \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i, \tag{5}$$

since $\sum_{i \in S_g} w_i \approx \sum_{i \in s_g} (M_g/m_g) w_i$ (see equation (A1) of the appendix). This implies

$$E_2[(t_2 - t_1)^2] \approx \text{Var}_2 \left\{ \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i \right\}$$

$$= \sum_{g=1}^{G} (M_g^2/[\{M_g - 1\} m_g])(1 - m_g/M_g)$$
$$* \left\{ \sum_{i \in S_g} (w_i r_i)^2 - \left(\sum_{i \in S_g} w_i r_i \right)^2 / M_g \right\}$$

$$\approx \sum_{g=1}^{G} ([M_g/m_g] - 1) \left\{ \sum_{i \in S_g} (w_i r_i)^2 \right\}. \tag{6}$$

Observe that equation (6) does *not* ignore the finite population corrections from the second phase of sampling.

3. THE JACKKNIFE VARIANCE ESTIMATOR

3.1 The Variance Estimator

We are now ready to discuss the jackknife. For $j \in F_h$, define the jackknife replicate $t_{(hj)2}$ as

$$t_{(hj)2} = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} \right\}, \tag{7}$$

where

$$w_{hji} = \begin{cases} w_i n_h / (n_h - 1) & \text{when } i \in U_{hj'} \text{ and } j' \neq j \\ 0 & \text{when } i \in U_{hj} \\ w_i & \text{when } i \in U_{h'j'} \text{ and } h' \neq h. \end{cases}$$

Similarly, we define

$$t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} y_i .$$

Following Rust (1985), the *jackknife variance estimator*, $v_{Jf} (f = 1 \text{ or } 2)$, is defined here simply as

$$v_{Jf} = \sum_{h=1}^{H} (n_h - 1)/n_h \sum_{j \in F_h} (t_{(hj)f} - t_f)^2. \tag{8}$$

This form is labeled $v_J^{(2)}$ in Krewski and Rao (1981, equation (2.4)). It is easy to show that $v_{J1} = v_{L1}$.

3.2 Why it Works (More Theory)

We will soon see that v_{J2} provides a nearly unbiased estimator for the variance of the reweighted expansion estimator in equation (2). Rao and Shao (1992) indirectly make the same claim (our equation (2) is the expectation of their estimator in Section 3.3, pp. 818-819). Their work, however, treats nonresponse as an additional phase of sample selection in which Poisson sampling (Särndal *et al.* 1992, p. 85) is used in place of stratified simple random sampling. Each first-phase sample element in the Rao and Shao (1992) setup is effectively a second-phase stratum. Consequently, the near unbiasedness of v_{J2} reduces to a special case of a result in Krewski and Rao (Rao and Shao 1992, p. 821).

What we have called the second-phase strata are reweighting classes in the Rao and Shao (1992) setup. Elements in the same class are assumed to have the same unknown probability of selection/response. *Conditional* on the realized subsample sizes within reweighting classes, Poisson sampling is equivalent to stratified simple random sampling. Rao and Shao's (1992) treatment, however, is *unconditional*.

Returning to the problem at hand, observe that

$$t_{(hj)2} - t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} \left\{ \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} - \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} r_{hji}}{\sum_{i \in s_g} w_{hji}} \right\},$$

where

$$r_{hji} = y_i - \sum_{k \in S_g} w_{hjk} y_k \Big/ \sum_{k \in S_g} w_{hjk} \quad \text{for} \quad i \in s_g.$$

Under mild conditions (see equations (A2) and (A3) in the appendix), we have the following analogue to equation (5):

$$t_{(hj)2} \approx t_{(hj)1} + \sum_{g=1}^{G} (M_g/m_g) \sum_{i \in s_g} w_{hji} r_{hji}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + [M_g/m_g] c_i r_{hji}), \tag{9}$$

where c_i is an indicator variable equal to 1 when i is in the subsample and zero otherwise.

Continuing,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + \{[M_g/m_g] c_i - 1\} r_{hji})$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_{hji}, \tag{10}$$

where $z_{hji} = y_i + \{[M_g/m_g] c_i - 1\} r_{hji}$. Again, since every m_g is large, it is not unreasonable to assume $r_{hji} \approx r_i$ (see equation (A4) in the appendix). Thus,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_i,$$

where $z_i = y_i + \{[M_g/m_g] c_i - 1\} r_i$. Using similar arguments, $t_2 \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_i z_i$. Since t_2 is linear in the z_i,

$$v_{J2} \approx v_{L1}\left(\sum \sum w_i z_i\right) = \sum_{h=1}^{H} (n_h/[n_h - 1])$$

$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i z_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right]^2 / n_h \right). \tag{11}$$

Let $e_i = M_g/m_g$ be the second-phase expansion factor for $i \in S_g$. Observe that c_i is a random variable with $E(c_i) = m_g/M_g$ and $E(c_i c_k) = (m_g/M_g)(m_g - 1)/(M_g - 1)$ for $i, k \in S_g$, $i \neq k$.

Now

$$E_2\left[\left(\sum_{i\in U_{hj}} w_i z_i\right)^2\right] \approx \left(\sum_{i\in U_{hj}} w_i y_i\right)^2 + \sum_{i\in U_{hj}} (e_i - 1)(w_i r_i)^2$$

$$- \sum_{g=1}^{G} \sum_{\substack{i,k\in S_g \cap U_{hj} \\ i\neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (12)$$

Similarly, letting F_h^* be the set of elements from selected clusters in the first-phase stratum h before subsampling, we have

$$E_2\left[\left(\sum_{j\in F_h} \sum_{i\in U_{hj}} w_i z_i\right)^2\right] = E_2\left[\left(\sum_{i\in F_h^*} w_i z_i\right)^2\right]$$

$$\approx \left(\sum_{i\in F_h^*} w_i y_i\right)^2 + \sum_{i\in F_h^*} (e_i - 1)(w_i r_i)^2$$

$$- \sum_{g=1}^{G} \sum_{\substack{i,k\in S_g \cap F_h^* \\ i\neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (13)$$

In the appendix, it is argued that under mild conditions that the last term in both equations (12) and (13) is negligible. As a result,

$$E_2(v_{J2}) \approx v_{J1} + \sum_{h=1}^{H} \sum_{i\in F_h^*} (e_i - 1)(w_i r_i)^2$$

$$= v_{J1} + \sum_{g=1}^{G} \sum_{i\in S_g} ([M_g/m_g] - 1)(w_i r_i)^2$$

$$\approx v_{L1} + E_2[(t_2 - t_1)^2], \quad (14)$$

which in turn implies that v_{J2} is a nearly unbiased estimator for $E[(t_2 - T)^2]$.

4. THE DOUBLE EXPANSION ESTIMATOR

An alternative to t_2, the *double expansion* estimator, has the form:

$$t_3 = \sum_{g=1}^{G} \sum_{i\in s_g} (M_g/m_g) w_i y_i. \quad (15)$$

The definition of a jackknife replicate for t_3 is unclear. One simple possibility is

$$t_{(hj)3} = \sum_{g=1}^{G} \sum_{i\in s_g} w_{hji}(M_g/m_g) y_i. \quad (16)$$

Another, perhaps more in the spirit of "replication", is

$$t^*_{(hj)3} = \sum_{g=1}^{G} \sum_{i\in s_g} w_{hji}(M_{ghj}/m_{ghj}) y_i, \quad (17)$$

where M_{ghj} is the number of elements in the first-phase sample (*i.e.*, in a cluster in the first-phase sample) that are in S_g but *not* U_{hj}. Similarly, m_{ghj} is the number of elements in the second-phase sample that are in s_g but *not* U_{hj}. Through counter-examples given in the appendix, we show that neither version of the replicate produces a jackknife variance estimator (v_{J3} from equation (8)) that is asymptotically unbiased in general.

5. A MONTE CARLO SIMULATION STUDY

5.1 Design of the Study

The results given so far in the text are asymptotic. In order to assess the accuracy of the jackknife as a variance estimator for the reweighted expansion estimator in a finite world, we undertook a Monte Carlo simulation study. At the same time, we assessed the accuracy of the two jackknife estimators suggested for the double expansion estimator in Section 4.

We used December 1990 Canadian Labour Force Survey (LFS) sample data for the province of Newfoundland to simulate a finite population, from which repeated samples were drawn. The LFS is the largest ongoing household sample survey conducted by Statistics Canada. Monthly data relating to the labour market is collected using a complex multi-stage sampling design with several levels of stratification. The details of the design of the survey prior to the 1991 redesign can be found in Singh, Drew, Gambino and Mayda (1990) and Stukel and Boyer (1992). In general, provinces are stratified into "economic regions", which are large areas of similar economic structure; Newfoundland has four such economic regions. The economic regions are further substratified into lower level substrata. The lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 clusters or *primary sampling units* (PSU's), which was an insufficient number from which to sample for the purposes of the simulation. Thus, the 45 strata were collapsed down to 18, each containing between 6 and 18 PSU's. In collapsing the strata, economic regions were kept intact, as were the Census Metropolitan Areas of St. John's and Cornerbrook.

For the Monte Carlo study, R = 4,000 samples were drawn from the Newfoundland "population" (which was 9,152 individuals), according to the following two-phase design: within each first-phase stratum, two PSU's were selected at the first phase using simple random sampling (SRS) *with* replacement. This yielded a total of 36 PSU's. All households within selected first-phase PSU's (as well as individuals within those households) were selected, resulting in a single-stage take-all cluster sample. At the second phase, all selected first-phase elements (individuals, treating each person in a PSU selected twice as two separate individuals) were restratified according to five age categories (< = 14, 15-24, 25-44, 45-64, > = 65), and second-phase sample elements (*i.e.*, individuals) were drawn using SRS *without* replacement sampling within each of the five second-phase strata.

We varied the second-phase stratum sample size to take on values m_g = 5, 10, 20, and 50 yielding overall second-phase sample sizes of m = 25, 50, 100, and 250. When the number of first-phase-sampled individuals in a second-phase stratum was less than our target m_g value, we planned to set $m_g = M_g$, but that event never occurred.

A popular rule of thumb for a "separate ratio estimator" such as the reweighted expansion estimator in equation (2) is that there should be at least 20 individuals within each second-phase stratum (see, for example, Särndal, Swensson and Wretman 1992, p. 270). By allowing m_g to be as small as 5 and 10, we are checking whether this rule is really necessary.

We considered two parameters of interest: T_y, the total number of employed, and T_y/T_z the employment rate. Here $T_y = \sum_{i \in U} y_i$, where $y_i = 1$ when individual i is employed; 0 otherwise. Similarly, $T_z = \sum_{i \in U} z_i$, where $z_i = 1$ when individual i is in the labour force (*i.e.*, either employed or unemployed); 0 otherwise. For each of the R = 4,000 samples, we calculated the reweighted expansion estimator (REE), t_2, given by equation (2), the double expansion estimator (DEE), t_3, given by equation (15), and the full first-phase expansion estimator (FFPE), t_1 given by equation (1). Although these estimators are defined for totals (applicable for total number of employed), it is a simple matter to extend them to ratios of totals (applicable for employment rate).

For each of the R = 4,000 second-phase samples, we calculated the jackknife variance corresponding to the reweighted expansion estimator and the double expansion estimator, given by equation (8) with $f = 2$ and $f = 3$ respectively. In the case of the double expansion estimator, we attempted both the replicates defined in equations (16) and (17), which we will refer to as variant 1 and 2, respectively.

For each of the R = 4,000 first-phase samples, we also calculated the jackknife variance corresponding to the full first-phase estimator for comparison purposes. This is given by equation (8) with $f = 1$.

For all of the above estimators and their corresponding jackknife variances, a number of frequentist properties were investigated. These are given below. For simplicity, they are expressed only in terms of estimates of the total number of employed.

The percent relative bias of the estimated number of employed with respect to the population value is estimated by

$$\mathrm{PRB}(t^*) = \{[E_M(t^*)/T_y] - 1\} \times 100, \qquad (18)$$

where

$$E_M(t^*) = (1/4{,}000) \sum_{r=1}^{4{,}000} t_r^*$$

is the Monte Carlo expectation of the point estimator t^* taken over the 4,000 samples. Here t^* can be either t_1, t_2, or t_3, and t_r^* is the value of t^* for sample r.

The percent relative bias of the jackknife variance estimator with respect to the true mean squared error is estimated by

$$\mathrm{PRB}[v_{Jf}(t^*)] =$$

$$(\{E_M[v_{Jf}(t^*)] - \mathrm{MSE}_{\mathrm{true}}\}/\mathrm{MSE}_{\mathrm{true}}) \times 100, \qquad (19)$$

where

$$E_M[v_{Jf}(t^*)] = (1/4{,}000) \sum_{r=1}^{4{,}000} v_{Jfr}(t^*),$$

$$\mathrm{MSE}_{\mathrm{true}} = (1/4{,}000) \sum_{r=1}^{4{,}000} (t_r^* - T_y)^2,$$

and $v_{Jfr}(t^*)$ is the value of $v_{Jf}(t^*)$ for sample r.

The (percent) coefficient of variation of the jackknife variance with respect to the true MSE is estimated by:

$$\mathrm{CV}[v_{Jf}(t^*)] =$$

$$(\{(1/4{,}000)\sum [v_{Jfr}(t^*) - \mathrm{MSE}_{\mathrm{true}}]^2\}^{½}/\mathrm{MSE}_{\mathrm{true}}) \times 100; \qquad (20)$$

that is, the estimated root mean squared error of the variance estimator divided by the estimated true MSE, expressed as a percentage.

5.2 Results of the Study

Table 1A gives the estimated percent relative biases of the three point estimates for the total number of employed using equation (18), and Table 1B gives the same for the employment rate. All biases are less than 1% in absolute value.

Table 1A
Percent Relative Bias of the Point Estimates for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	0.14	−0.3	−0.29	−0.56
DEE	–	0.16	−0.01	0.03	0.115
FFPE	0.04	–	–	–	–

Table 1B
Percent Relative Bias of the Point Estimates for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	−0.09	−0.31	−0.19	−0.26
DEE	–	−0.08	−0.27	−0.12	−0.13
FFPE	−0.09	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)

Not displayed are the Monte Carlo estimates of the mean squared errors (*i.e.*, the values of $\mathrm{MSE}_{\mathrm{true}}$) and the corresponding coefficients of variation from using either the reweighted or double expansion estimator. This is because the focus in this article is on mean squared error *estimation*. The mean squared errors (and coefficients of variation) from using the two estimators are comparable for each sample size (a relative difference in the coefficient of variation is roughly half of the corresponding relative difference in mean squared error). The reweighted expansion estimator is slightly more efficient when estimating the total number of employed individuals (*e.g.*, when $m_g = 5$, the double expansion estimator has 17% more mean squared error). There is less than a 1% difference in the mean squared errors from using the two approaches when estimating the employment rate. Not surprisingly, the mean squared errors for all estimators increase as the second-phase sample size decreases.

Table 2A gives the estimated percent relative biases of the jackknife variances for the total number of employed using equation (19), and Table 2B gives the same for the employment rate. Focusing first on Table 2A, the full first-phase estimator's variance is almost perfectly unbiased, at 0.94%. The jackknife for the reweighted expansion estimator works well, having small negative biases in the variances always less than −6%. The biases tend to become more negative (although not uniformly) as the second-phase sample sizes diminish.

Table 2A
Percent Relative Bias of Jackknife Variances
for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	−0.99	−2.51	−5.81	−5.13
DEE (Variant 1)	–	46.35	68.24	78.18	86.22
DEE (Variant 2)	–	101.59	278.44	654.99	1997.51
FFPE	0.94	–	–	–	–

Table 2B
Percent Relative Bias of Jackknife Variances
for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	−3.53	−3.45	−7.09	−6.55
DEE (Variant 1)	–	−2.46	−1.53	−5.21	−7.41
DEE (Variant 2)	–	−0.36	4.91	9.09	30.46
FFPE	2.08	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)
Variant 1 uses the jackknife replicates in equation (16)
Variant 2 uses the jackknife replicates in equation (17)

In contrast, both jackknife variants for the double expansion estimator fail miserably, with very large positive biases in the variances ranging from 46.35% to 1997.51%! The second variant is worse than the first, but both are well beyond the realm of acceptable behavior.

Table 2B repeats the analysis for the ratio estimate of employment rate. The results here are surprising since all variance estimators behave reasonably well, with the exception of variant 2 of the double expansion estimator when $m_g = 5$. Other than this case where the bias in the variance is 30.46%, all other biases are less than 10% in absolute value.

Overall, Table 2A and 2B provide strong support for using the jackknife variance estimator with a reweighted expansion estimator even when second-phase sample sizes are surprisingly small. By contrast, the jackknife can fail miserably for the double expansion estimator when estimating totals. Sometimes, however, variant 1 can also work reasonably well depending on the estimator and the data.

Although most studies focus on the *bias* of the variance estimators, it is also of secondary interest to look at the *coefficient of variation* of the variance estimators to see how stable the variance estimates themselves are. In Tables 3A and 3B, we investigate the estimated (percent) coefficients of variation corresponding to the total number of employed and the employment rate, respectively. In equation (20), the expression under the square root in the numerator gives the MSE of the variance, whose component parts are the square of the bias of the variance and the variance of the variance. For those entries in Tables 2A and 2B where the bias of the variance has been determined to be exceedingly large (say larger than 20%), the corresponding entries in Tables 3A and 3B are not reported (indicated by a *), since it is clear that those entries will be excessively large. In Table 3A, the estimated coefficients of variation corresponding to the reweighted expansion estimator range between 46.86% and 53.42%. Coefficients of variation of the magnitude exhibited here are typical for variance estimators, and have been encountered in other simulation studies relating to variances. See, for example, Kovačević and Yung (1997). To that end, note that even the estimated coefficients of variation corresponding to the full first-phase estimators are in the same range, and in fact, somewhat higher than those of the second-phase estimators in all cases.

Table 3B, which gives the coefficients of variation for the variances of the estimated employment rates, are entry by entry higher than their counterparts in Table 3A. In addition, all estimators exhibit the pattern that their corresponding coefficients of variation increase, quite substantially in fact, as the second-phase sample sizes diminish. This effect is more pronounced for the ratio estimators than it is for the estimators of the total. The very high coefficients of variation in the column $m_g = 5$ for both tables is not surprising, since the overall second-phase sample size (25) is actually smaller than the number of PSU's drawn in the first phase of sampling (36). In fact, a

Table 3A
Coefficient of Variation of Jackknife Variances for Total Number of Employed

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	51.33	49.3	46.86	53.42
DEE (Variant 1)	–	*	*	*	*
DEE (Variant 2)	–	*	*	*	*
FFPE	56.71	–	–	–	–

Table 3B
Coefficient of Variation of Jackknife Variances for Employment Rate

Estimator	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	–	59.28	65.66	74.26	103.06
DEE (Variant 1)	–	59.24	66.16	72.89	99.1
DEE (Variant 2)	–	60.94	73.2	92.71	*
FFPE	78.42	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)
Variant 1 uses the jackknife replicates in equation (16)
Variant 2 uses the jackknife replicates in equation (17)

more relevant realized sample count for the ratio estimator is the number of sampled individuals in the labour force (*i.e.*, in the denominator). This value varies from sample to sample and is often considerably less than 25.

6. EXTENDING THE REWEIGHTED EXPANSION ESTIMATOR

6.1 The Reweighted Expansion Estimator

It is not that difficult to develop a linearization variance estimator for the reweighted expansion estimator in equation (2). Suppose, however, one had a sample design with more than two phases or was interested in estimating the ratio of two totals. Linearization, although still possible, becomes increasingly cumbersome. The jackknife, on the other hand, does not.

It is a simple matter to generalize the results in Section 3 to p-phase sampling by induction. The h still refer the first-phase strata, but the g now denote the p-th-phase strata; S_g is the set of elements in the (p-1)th-phase sample from stratum g while s_g is the pth-phase subsample from g. The w_i in equation (2) are replaced with the a_i from (3) for the (p-1)th-phase estimator. Similarly, the $t_{(hj)2}$ in the jackknife are computed using a_{hji} from the (p-1)th phase in place of the w_{hji}.

It is also a simple matter (left to the reader) to replace the stratified cluster sample in the first phase of selection with a stratified multi-stage sample. The results in Section 3 follow as long as the first stage of the multi-stage sample is drawn with replacement.

Finally, it is not difficult to extend the results of Section 3 to more complicated estimators. Let U_2 be a vector of estimators each in the form of t_2 from equation (2). The mean squared error of any estimator $\Theta = g(U_2)$, where g is a smooth function, can be estimated with a jackknife in a nearly unbiased manner whenever the members of U_2 can be. This follows the proofs in the literature. Rao and Wu (1985), for example, address the asymptotic framework where the n_h are all bounded, while Wolter (1985; Chapter 4.5) treats the case where the n_h grow arbitrarily large.

6.2 Regression in the Second Phase

The estimator t_2 can be generalized into the regression estimator:

$$t_{2\text{reg}} = \sum_{i \in S} w_i \mathbf{x}_i \left(\sum_{i \in s} w_i e_i d_i \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\sum_{i \in s} w_i e_i d_i \mathbf{x}_i' y_i \right), \quad (21)$$

where S denotes the original sample, $\mathbf{x}_i$ is a row vector, d_i is a scalar, and there exists a row vector γ such that $d_i \gamma \mathbf{x}_i' = 1$ for all i. In practice, d_i is usually 1 for all i. A popular exception occurs when $\mathbf{x}_i = x_i$ and $d_i = 1/x_i$. In equation (2), $d_i = 1$ for all i, and $\mathbf{x}_i$ is a G-vector with a value of 1 in the g-th position and 0's elsewhere for $i \in S_g$.

Let

$$r_i = y_i - \mathbf{x}_i \left(\sum_{i \in S} w_i d_i \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\sum_{i \in S} w_i d_i \mathbf{x}_i' y_i \right).$$

The replicate $t_{2\text{reg}(hj)}$ has the same form as $t_{2\text{reg}}$ except that w_{hji} replaces w_i everywhere. Similarly, r_{hji} has the same form as r_i except that w_{hji} replaces w_i. Note that the e_i are unchanged from $t_{2\text{reg}}$ to $t_{2\text{reg}(hj)}$.

Since the sampling design hasn't changed, most of equation (6) stays as is except that now $(\sum_{i \in S_g} w_i r_i)^2$ is nonnegative rather than strictly zero. The interested reader can verify that equations (10) through (13) remain in their present form. It turns out that the jackknife has, if anything, an (approximate) upward bias in equation (14). That is to say, the jackknife is a *conservative* estimator of variance. Again, see the apppendix (equations (A6) through (A9)) for a formal statement of the asymptotic assumptions.

The bias in the jackknife disappears when $\sum_{i \in S_g} w_i r_i = 0$ for all g. Formally, this will happen when there exists G row vectors $\gamma_1, \ldots, \gamma_G$ such that $d_i \gamma_g \mathbf{x}_i' = 1$ when $i \in S_g$ and 0 otherwise (since $\sum_{i \in s} w_i r_i = \sum_{i \in s} d_i \gamma_g \mathbf{x}_i' w_i r_i = \gamma_g \sum_{i \in s} w_i d_i \mathbf{x}_i' r_i = \gamma_g \{\sum_{i \in s} w_i d_i \mathbf{x}_i' (y_i - \mathbf{x}_i [\sum_{i \in s} w_i d_i \mathbf{x}_i' \mathbf{x}_i]^{-1} \sum_{i \in s} w_i d_i \mathbf{x}_i' y_i)\} = 0$). When all $d_i = 1$, the existence of γ_g

means that either one member of x_i is an indicator variable equal to 1 when $i \in S_g$ and 0 otherwise, or one member of a linear transform of x_i is such an indicator variable.

7. CONCLUDING REMARKS

The main purpose of this paper was to show that a simple jackknife variance estimator can be nearly unbiased for an estimation strategy involving two-phase sampling as long as that strategy employs a reweighted expansion estimator and not a double expansion estimator. Since the theoretical results for the reweighted expansion estimator rely on asymptotic arguments, their practical application will depend on the context. Nevertheless, a Monte Carlo simulation study performed here suggests that the jackknife can be an effective estimator for the variance of a reweighted expansion estimator even with surprisingly small second-phase stratum sample sizes, that is, sizes of 5 and 10.

APPENDIX

The Design Consistency of the Reweighted Expansion Estimator

To establish the design consistency of t_2 in equation (2) it is sufficient to assume that the sample design and population values of the y_i are such that

$$\left\{\sum_{g=1}^{G}(M_g/m_g)\sum_{i\in s_g} w_i y_i/T\right\} - 1 = O_p(1/\sqrt{m}),$$

and, given *any* first-phase sample,

$$\left(\sum_{k\in S_g} w_k \Big/ \sum_{k\in s_g} w_k\right)(m_g/M_g) - 1 = O_p(1/\sqrt{m}) \qquad \text{(A1)}$$

for all g. These assumptions justify equation (5) in the text.

We assume in our analysis that G is bounded and that each m_g has the same asymptotic order as m. This is only possible when the S_g are determined *after* the first-phase sample has been drawn. Otherwise, the M_g would be random variables, and a minimum size for each m_g could not be guaranteed for all possible first-phase samples. In principle, we are assuming the existence of a mechanism for determining the S_g and the second-phase sampling fractions given any first-phase sample. By contrast, the exact values of G and the m_g can but need not be fixed before the first-phase sample is drawn.

A Comment on the Asymptotic Framework

Recall that the text showed that the jackknife contains a component that estimates the second-phase variance (*i.e.*, $E_2[(t_2 - t_1)^2]$) in an asymptotically unbiased manner given *any* first- phase sample (see equation (14)). As a result, that component also estimates the average (*i.e.*, unconditional) second-phase variance across all possible first-phase samples (*i.e.*, $E_1\{E_2[(t_2 - t_1)^2]\}$) in an asymptotically unbiased manner.

In our empirical work, we strayed from the sampling framework described above so that the results could be easily summarized. In particular, we defined the S_g beforehand, and let the M_g be random. When the first-phase sample was such that M_g was less than the desired m_g (say 50) in some second-phase stratum, we planned to choose all the individuals in S_g for the second-phase sample. As a result, there would be no contribution to the mean squared error (or bias) of t_2 from second-phase stratum g when that particular first-phase sample was selected, and so no asymptotic assumptions about m_g would be necessary. As it happened, in no simulation was M_g actually less than 50. Nevertheless, a decision rule about the second-phase sampling fractions was in place for every possible first-phase sample.

Jackknife Replicates

There are (at least) two distinct asymptotic frameworks for the first-phase sample. In the first, there is an arbitrarily large number of first-phase strata each of which is bounded in size; that is, each $1/n_h = O(1)$ while $1/H = O(1/m)$. In the second, all the first-phase strata are arbitrarily large; that is, $1/n_h = O(1/m)$. Under either framework, we assume that the number of elements in each cluster is O(1); that is to say, bounded.

Since every m_g is of the same asymptotic order as m, it is not unreasonable to assume under either regime that, given any first-phase sample,

$$\sum_{i\in S_g} w_{hji} \Big/ \sum_{i\in S_g} w_i - 1 = O_p(1/m), \qquad \text{(A2)}$$

and

$$\sum_{i\in s_g} w_{hji} \Big/ \sum_{i\in s_g} w_i - 1 = O_p(1/m), \qquad \text{(A3)}$$

which can be used to establish equation (9). Similarly, we assume that given any first-phase sample

$$\sum_{i\in S_g} w_{hji} y_i \Big/ \sum_{i\in S_g} w_i y_i - 1 = O_p(1/m), \qquad \text{(A4)}$$

which assures us that $r_{hji} - r_i = O_p(1/m)$.

Equations (12), (13), and (14)

Since the number of elements in each cluster is bounded, say by B. The third term on the right hand side of equation (12) has at most GB^2 terms, a bounded number.

Each of these terms is of order $1/m_g$ (formally, the probability that any one term is of asymptotic order greater than $1/m_g$ is zero). Consequently, the second line of equation (12) is asymptotically ignorable.

Equation (14) holds when each $1/n_h = O(1)$, because if each n_h is less than C (say), then the third term on the right hand side of equation (13) will be the sum of at most $G(BC)^2$ terms, a bounded number. Each of these terms is again of order $1/m_g$. Consequently, the second line of equation (13) is asymptotically ignorable.

Alternatively, suppose each $1/n_h$ were O($1/m$). We will assume that the sample design and population is such that, given any first-phase sample,

$$A_h = \sum_{i \in F_h^*} w_i(e_i c_i - 1) r_i \Big/ \sum_{i \in F_h^*} w_i y_i = O_p(1/\sqrt{m}) \quad \text{(A5)}$$

for all h. To see why this is a reasonable assumption, observe that conditioned on the first-phase sample, the denominator of A_h is a domain total – the sum of the $w_i y_i$ among the elements in F_h^*. Consequently, it is $O(m)$ (without loss of generality we can assume that all the w_i are $O(1)$). The numerator of A_h is the difference between an expansion estimator (the sum of the $w_i e_i c_i r_i$ in F_h^*) based on a stratified simple random sample and its target (the sum of the $w_i r_i$ in F_h^*). Equation (A.5) makes the modest assumption that the sampling design and population is such that this difference is $O_p(\sqrt{m})$ for every possible first-phase sample.

Under assumption (A5), $\sum_{i \in F_h^*} w_i z_i = \sum_{F_h^*} w_i y_i (1 + A_h)$ is approximately equal to $\sum_{i \in F_h^*} w_i y_i$, which implies $E_2[(\sum_{i \in F_h^*} w_i z_i)^2]/n_h \approx (\sum_{i \in F_h^*} w_i y_i)^2/n_h$. Equation (14) follows from this near equality and from equations (11) and (12) (since n_h is large, $n_h/(n_h - 1) \approx 1$).

Counter-examples to the Jackknifes for the Double Expansion Estimator

As a counter-example to the replicate form in equation (16), consider the situation where each cluster contains a single element, $H = G = 1$, and all the y_i values are equal to 1. As a result, $t_3 = T$, which means that t_3 has no variance. Unfortunately $t_{(1j)3} = T[n_1/(n_1 - 1)](m - 1)/m$ when $j \in s$ and $Tn_1/(n_1 - 1)$ otherwise. Thus, $(t_{(1j)3} - T)/T = O_p(1/m)$. Now v_{J3}/T^2 computed from the $t_{(1j)3}$ would also be $O(1/m)$ since it is the sum of n_1 terms of order $O(1/m^2)$.

Although v_{J3}/T^2 is $O(1/m)$, v_{J3} is not close enough to zero for our purposes. To see why, observe that if the y_i were all $N(1,1)$, then the relative variance of t_3 would be $1/m$, which is also $O(1/m)$. Thus, for v_{J3} to be nearly zero, v_{J3}/T^2 would have to be smaller than $O(1/m)$. It is not, and the jackknife variance estimator is not nearly unbiased.

As a counter-example to the replicate form in equation (17), consider the situation where each cluster is again a single element and all y_i values are equal to 1, but now $H = m$, $G = 1$, the population size in each h is N_0, $n_h = 2$ for all h, and $M_1 = 2m$. As a result, $T = t_3 = mN_0$, so that t_3 has no variance. The replicate $t^*_{(hj)3}$ can take on four possible values. If $hj \in s$ and $hj' \in s (j \neq j')$, then $t^*_{(hj)3} = [(m/2)(2m - 1)/(m - 1)]N_0$. If $hj \in s$ and $hj' \notin s$, then $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/(m - 1)]N_0$. If $hj \notin s$ and $hj' \in s$, then $t^*_{(hj)3} = [(m/2)(2m - 1)/m]N_0$. If $hj \notin s$ and $hj' \notin s$, then $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/m]N_0$. In all cases, $(t^*_{(hj)3} - T)/T = O_p(1/m)$, and so the jackknife variance estimator fails to be nearly unbiased.

The Two-phase Regression Estimator

To support the arguments in the text about the regression estimator in equation (21), we assume the sampling design and population values are such that the following asymptotic relationships hold. First,

$$\sum_{i \in S} w_i \mathbf{x}_i (\sum_{i \in s} w_i e_i d_i \mathbf{x}_i' \mathbf{x}_i)^{-1} d_i \mathbf{x}_i' - 1 = O_p(1/\sqrt{m}), \quad \text{(A6)}$$

which is a generalization of equation (A1). Likewise, equations (A2) and (A3) generalize to

$$\sum_{i \in S_g} w_{hji} d_i q_i \Big/ \sum_{i \in S_g} w_i d_i q_i - 1 = O_p(1/m), \quad \text{(A7)}$$

and

$$\sum_{i \in s_g} w_{hji} e_i d_i q_i \Big/ \sum_{i \in s_g} w_i e_i d_i q_i - 1 = O_p(1/m) \quad \text{(A8)}$$

for all q_i, where q_i is an element of the matrix $\mathbf{x}_i' \mathbf{x}_i$. Finally, the assumption in equation (A4) generalizes to

$$\sum_{i \in S_g} w_{hji} d_i p_i \Big/ \sum_{i \in S_g} w_i d_i p_i - 1 = O_p(1/m) \quad \text{(A9)}$$

for all p_i, where p_i is an element of the matrix $\mathbf{x}_i' y_i$.

REFERENCES

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOVAČEVIĆ, M.S., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization – an empirical study. *Survey Methodology*, 23, 1, 41-52.

KREWSKI, D., and RAO, J.N.K. (1981), Inferences from stratified samples: properties of linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. *Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4, 811-822.

RAO, J.N.K., and WU, C.F.J. (1985). Inferences from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey: 1984-1990*. Catalogue No. 71-526, Statistics Canada.

STUKEL, D.M., and BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD, 92-009E. Statistics Canada.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 91-98
Statistics Canada

A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France

GEORGES DECAUDIN and JEAN-CLAUDE LABAT[1]

ABSTRACT

Since France has no population registers, population censuses are the basis for its socio-demographic information system. However, between two censuses, some data must be updated, in particular at a high level of geographic detail, especially since censuses are tending, for various reasons, to be less frequent. In 1993, the Institut National de la Statistique et des Études Économiques (INSEE) set up a team whose objective was to propose a system to substantially improve the existing mechanism for making small area population estimates. Its task was twofold: to prepare an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is reported on here, is flexible and reliable, without being overly complex.

KEY WORDS: Population estimates; Administrative files; Robust estimation.

1. INTRODUCTION

In France, as in all countries that do not have population registers, censuses of the population are the cornerstone of the socio-demographic information system. However, censuses are quite massive operations that cannot at present be carried out more often than once every seven or eight years. In the interval between censuses, it is therefore necessary to update some information, especially at a high level of geographic detail, particularly since for various reasons, censuses are tending to be less frequent. Thus, small area population estimates are a major challenge for the Institut National de la Statistique et des Études Économiques (INSEE).

Despite the progress achieved in this field, the situation in 1993 still seemed fairly unsatisfactory. When figures from the 1990 population census were compared to the population estimates made on the basis of the previous census (1982) for the metropolitan departments, the differences noted were sometimes sizable.

INSEE therefore created a methodology team whose mission was to propose a system that would substantially improve the existing mechanism. Initially, the next census was to take place in 1997. It therefore seemed reasonable to have the new system operate on an experimental basis until the census, so as to see how well it worked before using it in actual production. When the census was postponed to 1999, it became more necessary to bring the project to a successful conclusion quickly, so as to be able to use the new system in 1996.

To achieve its objective, the team devoted itself, with maximum pragmatism, to a twofold task: to develop an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is described here, is not overly complex and seems effective. A more detailed description of it is provided in Decaudin and Labat (1996).

2. MAIN CONCLUSIONS

The team's main conclusions are as follows:

1) It is impossible to improve total population estimates using sample surveys, unless the survey is conducted on such a scale that it would be similar to a census.
2) No single administrative source adequately reflects changes in the population. At the local level, all sources can exhibit drift, breaks, jolts, *etc.*, which are not always easy to detect. Furthermore, even at the local level, it is often quite difficult if not impossible to get the agency responsible to provide explanatory details, much less corrections in the case of errors. In any event, it is unwise to rely on a single administrative source, however good it may be, since its permanency is never guaranteed.
3) On the other hand, total population estimates can be improved substantially by simultaneously using several sources. A "multi-source" system, similar to the one presented here but more rudimentary, was tested retrospectively over the intercensal period 1982-1990, for the 96 metropolitan departments. The mean error (mean deviation as an absolute value from the results of the March 1990 census) fell below 0.9%, whereas the mean error registered at the time, with the estimation system then in place, was 1.4%.

3. SIMULTANEOUS USE OF SEVERAL SOURCES

For using several sources jointly, different methods are possible.

A method that is universal – and easy to implement – is multiple regression. In simplified form, this amounts to using, for any area z, the following relationship:

$$P(n+1,z)/P(n,z) = c + \sum_S (k_S N_S(n+1,z)/N_S(n,z)),$$

[1] Georges Decaudin and Jean-Claude Labat, Institut National de la Statistique et des Études Économique, 18, Blvd. Adolphe-Pinard, 75765 Paris, CEDEX 14.

where $P(n, z)$ is the population of area z on January 1 of year n, the values $N_S(n, z)$ are the numbers from each source S on the same date and k_S are coefficients, which are estimated by multiple regression over a past period. Here c is a constant term that is used only in the regression, with calibration on the national population serving to correct any drift.

This method is used in various countries, including Canada and the United States (for example, see Statistics Canada 1987 and Long 1993). Nevertheless, it was not adopted because it has numerous drawbacks:

- it must be possible to estimate the coefficients, which requires data from each source extending back over a fairly long period;
- the coefficients can change over time, without it being possible to control this change;
- as noted above, the administrative sources are, for various reasons (changes in regulations, abrupt shifts in management, errors, *etc.*), subject to what might be called "anomalies". For each source S, the scope of these anomalies is reflected in part in the coefficient k_S, to an extent that depends on how great their medium-term effect has been over the calibration period [la période d'étalonnage]; but anomalies nevertheless occur in estimates with the same weight as the "good" data from the same source. The estimates are then highly distorted.

Another method is known as the "*composite*" method. Each source is used to estimate the population in one or more age classes: age class X, which is well-covered by the source, but also sometimes another class that definitely exhibits a pattern very similar to that of class X (for example, the "30-45" age group, if X represents the "under 18" age group). It is then necessary to have appropriate indicators for the other components of the population and correctly manage the consolidation of these estimates "in parts".

This type of method, used in the United States (Long 1993), seemed to us to be problematic, especially because of the difficulty of adequately dealing with "anomalies".

The *proposed "multi-source" system* is based on a robust synthesis of estimates from different sources. It combines demographic reasoning with purely statistical techniques. It draws on the experiments conducted by the INSEE's regional directorate in Brittany in the early 1970s (Laurent and Guéguen 1971; Guéguen 1972). Should one of the sources fail, such a system is not prevented from functioning, even though its performance may be somewhat diminished.

4. DEMOGRAPHIC BASE

The demographic reasoning which is at the base of the system is elementary: assuming that we know the total population $P(n)$ for an area on January 1 of year n, the population $P(n + 1)$ of the area on January 1 of year $n + 1$ is deduced by summing the two components of the change during year n: natural increase (births minus deaths), and net migration (immigrants minus emigrants).

$$P(n + 1) = P(n) + N(n) - D(n) + I(n) - E(n).$$

In France, natural increase data are provided annually at the commune level by vital statistics. If the latter are not yet available in final form, which is often the case in the third quarter of year $n + 1$, it is easy to estimate them with a low margin of uncertainty.

The only unknown, then, is net migration for year n: $SM(n) = I(n) - E(n)$ or what amounts to the same thing, the net migration rate $T(n) = SM(n)/P(n)$. In other words, estimating the population comes down to estimating net migration since the last date on which the population is known (or is assumed to be known), and vice versa.

In France, net migration figures are of some importance, although less so than in other countries such as Canada or the United States. In addition, they generally exhibit a certain inertia, at least at relatively aggregated geographic levels. One way to assess the influence of changes to them from one intercensal period to the next is to measure the errors that would have been committed during each period if the population had been estimated by using the average annual net migration rates for the preceding period. Over the period 1982-1990, for the departments (excluding Corsica), the mean end-of-period error (in 1990, at the end of eight years) would have been only 1.3%. It was not certain, when the team started its work, that much greater accuracy could be achieved. However, both in 1975 and in 1982, the mean error that would have been committed with the trend method would have been much greater: 2.8% and 2.7% respectively (over seven years). It would therefore seem that the period 1982-1990 was exceptional and that in the future the difference will again be more pronounced.

5. ESTIMATES FROM THE DIFFERENT SOURCES

From each source, using an appropriate method, we draw an estimate of annual net migration rate for the population as a whole. The methods that may be used depend on the data available.

For each of the sources tested and found to be "good", at least at the departmental level, a method is proposed. The five sources retained are the following: housing tax; electrical utility customers; children receiving family allowances; educational statistics; electoral file.

The data on the composition of households for tax purposes, which appear in the income tax files, are the sixth source that should provide very good results. However, to date, these data have been analysed for only a few departments, and the methodology for using them is not yet completely defined.

We also propose to integrate a trend estimate of the net migration rate into the system.

Two categories of methods are used. The first concerns the sources relating to households; the second concerns those relating to individuals.

5.1 Sources Relating to Households

Some sources provide information on changes in the number of households. This is the case with the files on *housing taxes* (HT) and *electrical utility customers* (EUC). The housing tax is one of the four main local direct taxes. As its name indicates, it applies to occupied dwellings, with main residences and secondary residences being treated separately. The housing tax file takes account of the situation on January 1 of the taxation year. Starting in the 1980s, the HT source was the basis for the departmental population estimates developed by INSEE (Descours 1992). In the early 1990s, it was replaced by the EUC source, in light of the distortions caused by a change to the HT management system which gradually worked its way through all departments.

The method adopted for using these sources follows classical principles. It leads directly to an estimate of the total population, and it involves three main stages:

1) estimating the number of households;
2) estimating average household size and from there, estimating the population of households;
3) adding the "non-household" population.

In the first stage, it is assumed that the number of households changes in accordance with the data supplied by the source (number of main residences for HT purposes or number of electrical utility customers). The second stage is more delicate. It is based on both the use of statistics on dependants from the HT files and on a trend estimate of average household size.

In the proposed "multi-source" system, we move on to the net migration rate, for comparison with other sources, using vital statistics data (*cf.* Section 4).

5.2 Sources Relating to Individuals

The other sources used concern individuals. Only a certain age group X of the population is generally covered adequately. The method then involves two main stages:

1) estimating, from the source, the net migration rate for the population aged X;
2) from there, estimating the net migration rate for the population as a whole.

The second stage is based on the following statistical relationship, observed in the past, between the change, from one period to another, of the overall net migration rate (T) and the change in the net migration rate for the population aged X (TX):

$$T_2 - T_1 = \delta_X (TX_2 - TX_1),$$

where δ_X is a coefficient close to 1, depending on the age group X. This relationship is similar to the one used by de Guibert-Lantoine (1987) to estimate the population on the basis of educational statistics.

For the corresponding age groups in the different sources used, the values, estimated by linear regression, of the coefficient δ_X(+/−2 standard deviations) are shown in tables 1 and 2.

Table 1
Estimates of δ_X on Departments, Excluding Corsica, Internal Net Migration

Period 1	Period 2	Age at end of period		
		0-19	10-14	35 and over
1962-1968	1968-1975	0.76 (+/- 0.04)	0.69 (+/-0.06)	1.24 (+/-0.09)
1968-1975	1975-1982	0.77 (+/-0.03)	0.88 (+/-0.06)	1.56 (+/-0.08)
1975-1982	1982-1990	0.70 (+/-0.11)	0.49 (+/-0.10)	1.26 (+/-0.17)

Table 2
Estimates of δ_X Over the Two Periods 1975-1982 and 1982-1990, Excluding Corsica, Total Net Migration

	Age at end of period		
	0-18	9-15	35 and over
Departments	0.65 (+/-0.11)	0.57 (+/-0.10)	1.22 (+/-0.16)
Department – employment zone	0.65 (+/-0.04)	0.59 (+/-0.04)	1.17 (+/-0.06)

The approach followed in the first stage depends on the source:

Electoral File

Annual migration figures for voters in the selected age group (30 and over) are supplied directly by the electoral file managed by INSEE. We go from the rate of net migration of voters to the residential net migration rate by dividing the former by a coefficient reflecting the magnitude of the change in the electoral file.

Educational Statistics

The net migration figure for those in the 5-9 age group is obtained by subtracting their number in year n from that of the same cohorts the next year (that is, from those in the 6-10 age group in year $n + 1$) and deducting deaths.

Children Receiving Family Allowances

The number of persons in the 0-17 age group is estimated on the assumption that it evolves similarly to the number of children receiving family allowances. From this a figure for the net migration of young persons is obtained by comparing this estimate to a hypothetical change in the youth population without migration, that is, a change due solely to natural increase.

6. SYNTHESIS

6.1 Principles

The different basic estimates of the annual net migration rate are treated statistically in order to obtain a "synthetic rate", to be used as the final estimate. The treatment serves to eliminate outliers, underweight suspect values and, more generally, assign to each source a weight that reflects its performance.

More specifically, since each source can "drift", the different basic estimates are generally biased; they are first corrected for the national bias of the corresponding source for the year considered, a bias that is estimated in advance. In proceeding in this way, we implicitly assume that the difference between the local bias and the national bias is minor in relation to the irreducible unexplained portion of the difference (flou irréductible). Once we have estimates for a number of years, it should be possible to test this hypothesis and if necessary, replace it with one that corresponds more closely to reality, so as to improve the correction of biases at the local level.

It should be noted that such a seemingly simple operation as correcting the national bias nevertheless requires several precautions. The solution that consists in carrying out a gross calibration on the national net migration rate, considered by definition as a good reference, is not very satisfactory, owing to anomalies that may distort the calibration. It is therefore preferable to estimate the biases by means of a process in which we also eliminate anomalies. The process is similar to the one used for synthesis, which is described below. However, the determination of biases, assumed to be national in scope and therefore calculated for 96 departments, is less sensitive to anomalies than the determination of synthetic rates, calculated over a small number of sources. Only major anomalies are likely to significantly throw off the calibration of the rates and must therefore be corrected.

The "synthetic" net migration rate is a weighted mean of the basic estimates thus calibrated. Each source S is assigned an initial weight W_S that is supposed to reflect its medium-term accuracy. But in addition, for a given year and area, this weight is modulated to take account of the plausibility of the corresponding rate. Thus, if a rate is "abnormally distant" from the rates obtained from other sources – in practice, from a central value for all rates for the area – its weight is cancelled or reduced. For this, we look at the distance between the rate obtained from each source and the central value identified, and we compare it to a "norm" of distance NO_S specific to the source, determined empirically on the basis of the data available: if the distance is less than "a times the norm", the weight is not automatically changed; if it is greater than "b times the norm", it is set at 0; between the two, the weight is multiplied by a coefficient, included between 0 and 1, calculated by interpolation.

Note that the trend estimate is formally treated like those from exogenous sources; its weight is cancelled when it is considered as implausible because it is too far from the other estimates.

The synthesis is achieved automatically, which ensures homogeneity and an explicit logic to the treatments carried out. This does not, however, eliminate the need to control the results obtained.

6.2 Theoretical Presentation

On the theoretical level, we sought to use reasonings and robust estimation techniques, such as described in Hoaglin, Mosteller and Tukey (1983). The method adopted falls within the framework of M-estimators of central tendency and more specifically in the category of W-estimators, which use the reweighted least squares algorithm.

Since the net migration rates for year n and area z obtained from different sources S (and corrected for their national biases) are denoted $TC_S(n, z)$, the synthetic rate $T(n, z)$ solves the implicit equation:

$$\sum_S W_S \cdot NO_S \cdot \Psi\left(\frac{TC_s(n,z) - T(n,z)}{NO_S}\right) = 0,$$

where the function Ψ is of the type that redescends to a finite rejection point:

$$\Psi(r) = r \qquad \text{for } |r| \le a,$$

$$\Psi(r) = r\frac{b-|r|}{b-a} \qquad \text{for } a < |r| \le b,$$

$$\Psi(r) = 0 \qquad \text{otherwise.}$$

Using an iterative process, we can gradually refine the automatic processing of suspect data.

6.3 First Analysis of the Distances From Each Rate to the Central Value for the Rates

1) For each area z we calculate a first central value of the "calibrated" rates $TC_S(n, z)$. The central value used must not be overly sensitive to the possible existence of quite distant values for some sources, but at the same time it must be influenced by a source to the extent that the source is on average more accurate. Under these conditions, rather than choosing the median – which would meet the first condition – we use a statistic of rank that is a little more elaborate but nevertheless simple, owing to the small number of values; this statistic is the mean, weighted by respectively 1/2, 1/4, 1/4, of the three quartiles:
   - the median of the rates $TC_S(n, z)$ weighted by the initial weights W_S,
   - the lower quartile (Q1) of the weighted rates,
   - the upper quartile (Q3) of the weighted rates.

2) The rates $T1(n, z)$ thus obtained are calibrated on the net migration rate for the higher level, by simple translation:

$$TC1(n,z) = T1(n,z) + TREF(n) - \sum_z (T1(n,z)P(n,z)) \Big/ \sum_z P(n,z)$$

where $P(n, z)$ is the population of area z on January 1 of year n and $TREF(n)$ is the net migration rate for the higher level (the national rate for the departmental synthesis).

3) For each area, we calculate the differences between each rate and this calibrated central value:

$$EC1_S(n, z) = | TC_S(n, z) - TC1(n, z) |.$$

4) For each source and each area, the size of this difference is assessed in relation to the "norm" of distance NO_S specific to the source. This "norm" is determined empirically on the basis of the available data: theoretically it is the average of the distances observed in the past, excluding anomalies. The result is a first modulation of the weight originally assigned to this source:
   - if $EC1_S(n, z) \leq a1\, NO_S$, where $a1$ is a parameter to be chosen (in the vicinity of 2), we do not change W_S, the initial weight for S. In other words, if $WM1_S(n, z)$ is the modulation coefficient of W_S (coefficient included between 0 and 1), we take $WM1_S(n, z) = 1$;
   - if $EC1_S(n, z) > b1\, NO_S$, where $b1$ is another parameter (in the vicinity of 3), we set W_S at 0, meaning that we eliminate source S: $WM1_S(n, z) = 0$;
   - if $a1\, NO_S < EC1_S(n, z) \leq b1\, NO_S$, we interpolate $WM1_S(n, z)$ as a function of the value of $EC1_S(n, z)$:

$$WM1_S(n, z) = (b1\, NO_S - EC1_S(n, z)) / ((b1 - a1) NO_S).$$

5) At the end of this first phase, we therefore have new weights specific to each source and each area, which would allow us to locally eliminate or underweight suspect rates: $W1_S(n, z) = W_S WM1_S(n, z)$.

6.4 Iterations

1) Using the weights thus modified $W1_S(n, z)$, we estimate a new central value for each area, this time taking the weighted average of the rates:

$$T2(n, z) = \sum_S (TC_S(n, z) W1_S(n, z)) \Big/ \sum_S W1_S(n, z).$$

2) We calibrate each rate $T2(n, z)$ on the net migration rate for the higher level, by translation. We obtain $TC2(n, z)$.
3) We calculate, in each area, the differences between each rate and the calibrated average rate: $EC2_S(n, z) = | TC_S(n, z) - TC2(n, z) |$. Using these differences, we calculate new modulation coefficients for the initial weights, using the parameters $a2$ and $b2$, which may be different from $a1$ and $b1$ (theoretically they would be lower). We thus obtain new weights $W2_S(n, z)$ which more effectively take account of anomalies, since the latter are assessed in relation to a better central tendency. With these weights, we estimate a new synthetic rate $T3(n, z)$, which is calibrated on the higher level to obtain $TC3(n, z)$.
4) The operations described in point 3 are repeated with the same parameters $a2$ and $b2$. The tests conducted at the departmental level over the period 1982-1990 show that the convergence is generally rapid; the rates are quite often stabilized by the fourth iteration.

7. IMPLEMENTATION AT THE DEPARTMENTAL LEVEL

The estimation system outlined above, which is operationalized for 1990 and subsequent years, was implemented by the project team for the year 1990 at the departmental level, with the following five sources: housing tax (HT), electrical utility customers (EUC), family allowances (FA), educational statistics (ES), electoral file (EF), plus the trend estimate (TREND).

Figure 1 shows the results obtained for several departments. Table 3 shows the values of the weights and norms used to make the system operate. This table also shows certain statistics obtained from the synthesis of the net migration rates; in particular they concern the differences between the rates obtained from each source and the synthetic rates.

Table 3
Implementation for Year 1990 at Department Level
Parameters and Statistics

	HT	EUC	FA	ES	EF	TEND
Weight	115	100	80	70	80	100
Norm	0.15	0.17	0.19	0.20	0.19	0.12
Number of rates	96	96	89	96	94	96
Average distance	0.55	0.14	0.30	0.19	0.14	0.13
Number of "aberrant" rates	37	2	17	3	1	6
Average of distances without "aberrant" rates	0.15	0.13	0.16	0.16	0.13	0.11

Note: - Coefficients (a; b) applied to norms: (2,5; 3,5) in the first iteration, then (2; 3).
- The values of the distances and norms correspond to rates expressed as a %.
- Distances are calculated in relation to the synthetic rates after three iterations.
- "Aberrant" rates are those for which the weight is cancelled after three iterations.

The results suggest that the system is even more effective than indicated by the summary retrospective test carried out on the 1982-1990 intercensal period with the same sources. Aside from the HT source, which is still distorted, the estimates from the different sources are more convergent than they were on average in the retrospective test (see Table 4).

There is nothing surprising about this, given the rudimentary state of the system tested on the 1982-1990 intercensal period. The data used were rough or even

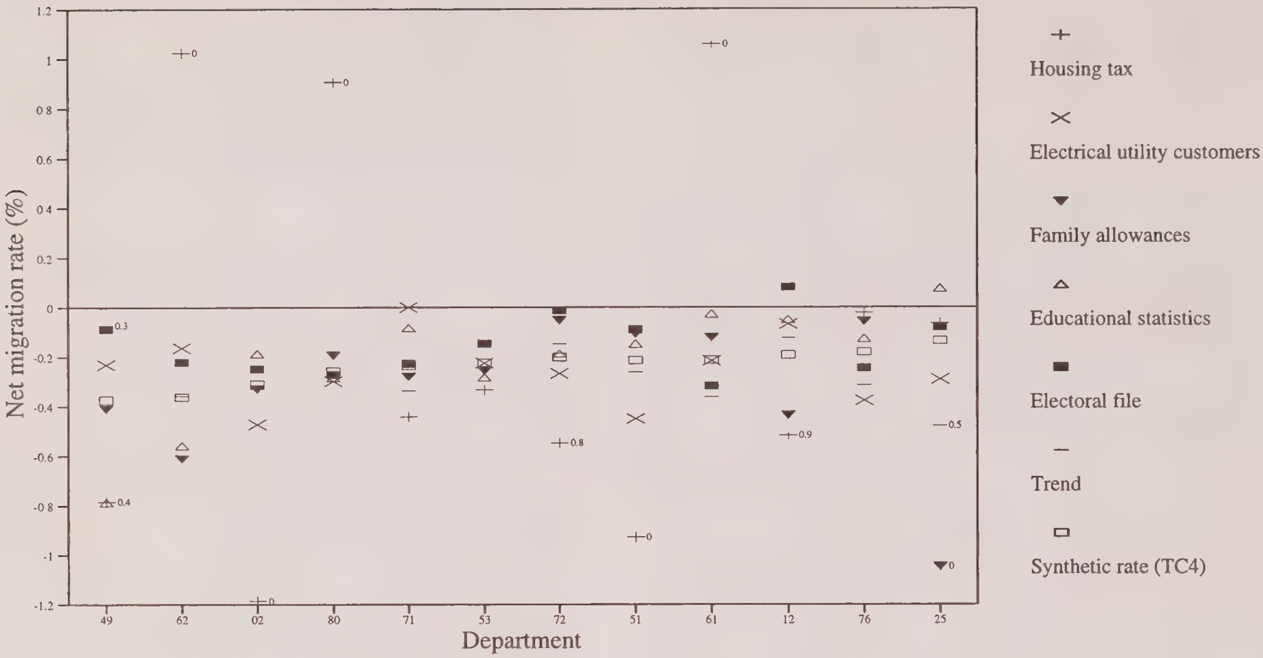


Figure 1: Summary of Net Migration Rates for 1990 for Twelve Departments, Identified by Number (49, 62, *etc.*).
Note: TC4 is the synthetic rate obtained after three iterations. Where the weight for a source has been eliminated or reduced, the value of the modulation coefficient (WM3) is shown.

fragmentary, owing to the difficulty of assembling, in 1993, management data for years past (1982, …); in addition, the relationships used to draw an estimate of the net migration rate from each source were simplistic; and lastly, the method of synthesis was less elaborate.

It should be noted that the integration of other sources – income tax data in particular – can only further reinforce the effectiveness of the system.

Table 4

Mean of Distance in Retrospective Test

	TH	EDF	AF	EN	FE
1982	0.26	0.34	0.50	0.47	0.34
1983	0.28	0.33	0.48	0.47	0.32
1984	0.23	0.28	0.40	0.45	0.34
1985	0.24	0.31	0.48	0.44	0.32
1986	0.23	0.33	0.40	0.33	
1987	0.40	0.28	0.41	0.27	
1988	0.84	0.29	0.30	0.37	0.24
1989	0.97	0.21	0.30	0.33	0.35
Overall mean	0.43	0.30	0.41	0.39	0.32

Notes: -The number of rates per year is generally 96, except for FA (89) and EF (94).
-The "electoral file" source did not provide rates for 1986 or 1987.
-The "housing tax" source began to be distorted in 1987.
-The values of the differences correspond to rates expressed as a %.

8. SUPPLEMENTS

8.1 Sub-Departmental Levels

The use of some sources may become risky at a geographic level below the departmental level. There are various reasons for this: because the hypotheses on which the method is based become fragile, because the numbers are small, *etc.* This is especially the case with educational statistics.

However, it should be possible to operate the system for employment areas, or more specifically for cross-tabulations of department and employment area (there are approximately 420 such areas), which serve to ensure consistency with the departmental level. This should not involve too many risks, for the following reasons:

- a certain deterioration of performance in relation to the departmental estimates is acceptable, especially since the departmental estimates should be of good quality;
- the data from the income tax files should be quite useful;
- trend estimation and calibration on estimates at higher geographic levels (in this case the departmental estimates) both act as safeguards.

Of course, there is nothing prohibiting the use of the system to produce estimates for other sub-departmental geographic units.

At the departmental level, it does not seem useful to adapt the parameters (initial weights and norms) to population size; on the other hand, for sub-departmental

levels, such an adaptation appears essential. Otherwise we run the risk of being much too strict for small areas. It would seem that a norm function of the following type might be appropriate:

$$NO_S = \alpha P^{\beta},$$

where NO_S is the norm for source S, P is the population of the area and α and β are two parameters that hypothetically depend on source S. The parameter β is obviously negative. If β equals -0.25, the norm doubles when the population is divided by 16. It also appears that the type of geographic area has an effect: the unexplained portion (le flou) would on average be greater for a commune of 50,000 inhabitants than for an employment area of the same size. The parameters α and β must be defined for each sub-departmental source, and where applicable, for each type of area.

8.2 Timetable

The greater the number of sources, the better the system functions. However, for a given year, data from the different sources become available at different times. Since the system is able to function with a variable number of sources, one can develop, at least at the departmental level, several sets of estimates for January 1 of year n: for example, interim estimates in the third quarter of year n, based on the first sources available, then semi-definitive estimates in the third quarter of year $n + 1$, based on more sources, and then final estimates in the third quarter of year $n + 2$. Different factors must be taken into account: the complexity of an operation, and the magnitude of the changes due to the addition of a source. It will be possible to assess the latter factor by simulations on the first years of implementation of the system.

8.3 Integration of an Additional Source

The system is flexible and modular. Therefore, integrating a new source into it does not pose any particular problem. It is merely a matter of determining the method to be used in order to obtain a good estimate of the net migration rate for each area. The range of methods envisaged by the team is large enough that in most cases, it should be possible to find a type of method that is appropriate to the source.

To determine the parameters (initial weight and norm) to be assigned to the new source in the synthesis, we suggest putting the system through a dry run, with parameters set arbitrarily but reasonably; it is obviously wise to start with a fairly high norm and a fairly low weight. By analysing the differences obtained between the net migration rates obtained from the new source and the synthetic rates, a better norm can be determined. The weight can then be adapted accordingly, using (for lack of anything better) an assumed relationship of quasi-proportionality between the weight and the inverse of the square of the norm. Obviously, this process can be iterated, with the parameters of the other sources also being changed as required. However, the tests conducted at the departmental level on the period 1982-1990 appear to show that the overall performance of the system is not highly sensitive to changes – even sizable ones – in the initial weights; it is therefore not necessary to determine these weights with great precision – nor, indeed, is it possible to do so – before the next census.

9. CONCLUSION

The "multi-source" population estimation system presented here is robust and flexible, without being overly complex. It can function with a variable number of sources. To integrate a new source into it, no long historical observation period is required. Aberrant data are detected automatically and corrected, so that they do not distort the estimates. The experiments carried out, while still not numerous, indicate that this system is effective. After a debugging and break-in period, it should be possible to use the system in production without too many risks pending the results of the next population census, planned for 1999.

ACKNOWLEDGEMENTS

This article results from the thinking and efforts of a team, led by the authors, which consisted of: Xavier Berne, Michel David, Michel De Bie, Sophie Destandau, Jacques Leclercq, Françoise Lemoine, Catherine Marquis and Marc Simon. The team benefited from the assistance of several departments of INSEE. The Statistical Methods Unit and its chief, Jean-Claude Deville, deserve special mention. The authors also wish to thank Philippe Ravalet for his contribution to the theoretical aspect of this article, as well as the editorial staff of *Survey Methodology* and the members of the editorial jury for their constructive comments.

REFERENCES

DECAUDIN, G., and LABAT, J.-C. (1996). Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population. Document de travail de méthodologie statistique, n° 9601. INSEE. Paris.

DESCOURS, L. (1992). Estimation de populations locales par la méthode de la taxe d'habitation. *Actes des Journées de méthodologie statistique,* 13 and 14 March 1991. INSEE. Paris.

GUÉGUEN, Y. (1972). Estimation de la population des villes bretonnes au 1.1.1971. *Sextant*, n° 4. INSEE. Rennes.

de GUIBERT-LANTOINE, C. (1987). Estimations de population par département en France entre deux recensements. *Population*, 6, 881-910.

HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

LAURENT, L., and GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE. Rennes.

LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue No. 91-528E. Ottawa.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 99-107
Statistics Canada

An Adaptive Procedure for the Robust Estimation of the Rate of Change of Investment

PHILIPPE RAVALET[1]

ABSTRACT

The presence of outliers in survey data is a recurring problem in applied statistics, and the INSEE survey on industrial investment is not immune from this. The forecasting of the rate of growth of capital investment expenditures in industry therefore comes down to robust estimation of a total in a finite population. The first part of this article analyses the estimator currently used in the Investment Survey. We show that it follows a strategy of reweighting the linear estimator. But the strict dichotomy imposed between outliers – all assumed to be nonrepresentative – and other points is not fully satisfactory from either a theoretical or a practical standpoint. These flaws can be overcome by adopting a model-based approach and estimating by GM-estimators, applied to the case of a finite population. We then construct a robust adaptive procedure that determines the appropriate estimator on the basis of the residuals observed in the sample in cases where the residuals may be assumed to be symmetrical. Lastly, this method is applied to the data from the Investment Survey for the period 1990-1995.

KEY WORDS: Economic surveys; Outliers; Robust estimation; GM estimator; Adaptive procedure.

1. INTRODUCTION

Since 1952, the Institut National de la Statistique et des Études Économiques (INSEE) has been conducting an investment survey that provides estimates of the future trend of capital investment expenditures in industry, well before the National Accounts are released or the findings of exhaustive surveys are published. The estimation of the rate of investment growth is based on the declarations of some 2,500 company heads concerning their intentions to purchase capital goods.

The almost systematic presence of outliers in these data is a major problem. Outliers can seriously distort the estimate of the average growth rate and lead to unacceptable results. According to Chambers (1986), two types of outliers may be distinguished. Nonrepresentative points designate either measurement errors, which survey staff strive to correct during data collection, or unique individuals in the population. By contrast, representative outliers designate individuals which, while somewhat unusual, cannot be considered exceptional. There are undoubtedly similar individuals in the population not questioned, and the information that they contain must be integrated into the estimate.

The problem posed here is that of robust estimation of a total in a finite population with auxiliary information, a problem to which theory provides no definitive answer. Nevertheless, various techniques, reviewed in Lee (1995), can be applied. The estimation method currently used in the Investment Survey follows the logic of reweighting the linear estimator, following Hidiroglou and Srinath (1981). However, the identification and treatment of outliers are not entirely satisfactory. In particular, all outliers are assumed to be nonrepresentative, and the dichotomy between "normal" points and outliers makes the estimation quite sensitive to the choice of outliers.

The introduction of a linear superpopulation model, which describes the change in investment at the level of individuals, enables us to better assess the unusual nature of an observation and determine how representative it is. Its estimation by means of GM-estimators is then an attractive alternative to the least squares method, whose absence of bias is quite costly in terms of variance. The adjustment of the weight function depends at the outset on characteristics of the population according to criteria now well described in the literature. Since these characteristics can change not only from one stratum to another but also over time, the significance of an adaptive procedure is obvious. On the basis of a first robust estimate, we determine the appearance of the distribution of residuals, and then we choose the estimator to be used according to a predefined rule. Following Hogg, Bril, Han and Yul (1988), we construct an adaptive procedure based on indicators of tail weight and concentration estimated from the sample, since the residuals are not expected to be asymmetrical. This procedure is applied to the data from the Investment Survey for the period 1990-1995.

2. ESTIMATOR FOR THE INVESTMENT SURVEY

2.1 Estimation Principle

In a finite population $U = \{1, ..., N\}$, which here represents a stratum of the survey, a sample $s = \{1, ..., n\}$ of size n, is drawn, and $\bar{s} = \{n+1, ..., N\}$ designates the population not questioned. Each company is questioned on

[1] Philippe Ravalet, Division des enquêtes de conjoncture, INSEE, 15 Bd. G. Péri, BP 100, 92244 MALAKOFF CEDEX.

its investment expenditures for two consecutive years $t-1$ and t, denoted respectively x and y.

Knowing the total amount X of investments for year $t-1$ in the population, we can deduce from the estimate $\hat{Y}$ of total investments for year t the average rate of change of equipment expenditures between $t-1$ and t:

$$\hat{\theta} = \frac{\hat{Y} - X}{X}.$$

To simplify the notations, we define the parameter $\Theta = 1 + \theta = Y/X$, estimated by $\hat{\Theta} = \hat{Y}/X$.

The estimator currently used in the INSEE survey draws on the ratio method, with the level of investment in $t-1$ as auxiliary information:

$$\hat{Y}_{\text{ratio}} = \frac{X}{\sum_s x_i} \sum_s y_i.$$

This estimator may be written as a weighted linear estimator:

$$\hat{Y}_{\text{ratio}} = \sum_s w_i z_i. \qquad (1)$$

In this expression, $w_i = Xx_i / \sum_s x_j$ is the weight of individual i and $z_i = y_i / x_i$ is the annual change in its investment. Such an estimator will be sensitive to the presence of outliers on both z and w. *An atypical point* will exhibit a change z that is very different from that of the others, while an *influential point* will have a weight w that is large enough to attract, by leverage, the average rate of change of the stratum towards its own rate of change. Since the decisive criterion for characterizing an observation as an outlier is that the product wz is large enough to distort the estimate $\hat{Y}_{\text{ratio}}$, the distinction between atypical points and influential points is, of course, arbitrary. The generic term *large investors* (or LI for short) will designate these outliers as a group, while the term *extrapolatables* will refer to the other individuals in the sample.

Having carried out an *a posteriori* partition of the sample s = {LI} ∪ {*extrapolatables*}, we estimate the total investments of the rest of the population $\bar{s}$ on the basis of the behaviour of only the extrapolatable individuals according to the ratio method:

$$\hat{Y}_{\text{LI}} = \sum_s y_i + \left(\sum_{\bar{s}} x_i\right) \frac{\sum_{\{\text{extra}\}} y_i}{\sum_{\{\text{extra}\}} x_i}. \qquad (2)$$

In (2), the weight of the extrapolatables $1 + \sum_{\bar{s}} x_i / \sum_{\{\text{extra}\}} x_i$ is quite strictly greater than the weight of the large investors, which is equal to 1.

2.2 Selection of Large Investors

The large investors are selected within each stratum on the basis of their influence on the estimation of Θ according to an iterative procedure. At the outset, all individuals are assumed to be extrapolatable, and for each of them we calculate a not-taken-into-account index, measuring the impact on $\hat{\Theta}$ of its exclusion from the sample, NTIA = $(\hat{Y}^i_{\text{LI}} - \hat{Y}_{\text{LI}})/X$ where $\hat{Y}^i_{\text{LI}}$ is the estimated total without individual i.

The firm with the largest NTIA index in absolute value is said to be a large investor. $\hat{Y}_{\text{LI}}$ is then re-estimated with this new partition of U, and then the next large investor is identified. The selection stops when all extrapolatable individuals' have an influence on the estimate that is below a given threshold. The greater the number and mass of observations, the easier it is to verify this condition. Conversely, it will prove impossible to verify the condition if the number of individuals is too small; in that case, the survey manager merely makes sure that no individual has a much greater influence than the others, thus introducing an element of subjectivity into the procedure.

By this iterative mechanism, the usual phases of detection and treatment of outliers are carried out simultaneously. The main problem is that the status of an individual is not an intrinsic characteristic but instead depends on the composition of the sample. This can change from one survey to another. In addition, in certain hypothetical cases (Ravalet 1996), this procedure can lead to the unnecessary exclusion of some individuals, since at no point is the status of large investor called into question.

2.3 Strategy for Reweighting the Linear Estimator

The estimator LI in fact follows from the strategy for reweighting the linear estimator (1) presented by Hidiroglou and Srinath (1981) using the example of estimation of a total without auxiliary information. Having already carried out a partition $s = s_1 \cup s_2$ of the sample distinguishing the outliers s_1 (numbering n_1) from the other observations s_2, the authors propose to reduce, in $\hat{Y} = (N/n)\sum_s y_i$, the weight N/n of the outliers to a lower value λ by positing

$$\hat{Y}_\lambda = \lambda \sum_{s_1} y_i + \frac{N - \lambda n_1}{n - n_1} \sum_{s_2} y_i$$

and

$$\hat{Y}_\lambda = \sum_s y_i + \frac{N-n}{n-n_1} \sum_{s_2} y_i + n_1(\lambda - 1)\left[\frac{1}{n_1}\sum_{s_1} y_i - \frac{1}{n-n_1}\sum_{s_2} y_i\right].$$

The optimal value of λ that minimizes the mean square deviation of this estimator, whether or not conditional on the number of outliers in the sample, depends on several parameters of the population. Without prior information, the choice of λ is a delicate one.

Applied to the case of the estimator of the ratio with auxiliary variable x, this is written as:

$$\hat{Y}_{\text{ratio}\,\lambda} = \sum_{s} y_i + \sum_{\bar{s}} x_i \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} + (\lambda - 1)\left(\frac{\sum_{s_1} y_i}{\sum_{s_1} x_i} - \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i}\right)\sum_{s_1} x_i. \quad (3)$$

The first two terms of the second member of (3) form an estimate of the total Y, under the implicit hypothesis that all outliers are in the sample, and the third is a correction taking account of the possible presence of outliers in the population not questioned. This correction is a function of the λ selected and the difference in average behaviour between the two types of individuals estimated in the sample.

When (2) and (3) are considered together, it may be seen that the estimator LI is formally equivalent to the case $\lambda = 1$. The use of $\hat{Y}_{\text{LI}}$ thus implicitly assumes that the outliers have been correctly identified and are all non-representative. In Ravalet (1996), it was shown that these two hypotheses were unfortunately seldom verified in the context of the Investment Survey.

Since the identification procedure is manual and the criterion used is relatively *ad hoc* in the absence of any hypothesis on the population, it is not impossible that some outliers will escape selection. The use of the ratio on the extrapolatables then poses the problem of the robustness of the estimation in relation to the choice of large investors. In addition, it is unlikely that all these points are unique. The atypical points, which are especially numerous among small and medium-sized firms, should instead be considered as representative. However, choosing $\lambda > 1$ would inevitably raise the question of the robustness of the third term of (3).

To try to compensate for these defects, changes to the estimator $\hat{Y}_{\text{LI}}$ are possible. For example, the mean of the extrapolatables may be replaced by a more robust estimator, and only the nonrepresentative points are designated as large investors. This technique fits into the more general framework of M-estimators, in which the existence of a model facilitates both the detection and treatment of outliers (Lee 1995). It is then no longer a matter of constructing a strict dichotomy between outliers and other points but rather of defining areas of varying representativeness.

3. ROBUST ESTIMATION BY GM-ESTIMATORS

3.1 The Linear Model and GM-Estimators

Assume the existence of a linear model ξ that links together, for the overall population U, investments x and y on dates $t-1$ and t.

$$\xi : y_i = \beta x_i + \epsilon_i$$

with

$$E(\epsilon_i) = 0$$
$$E(\epsilon_i \epsilon_j) = 0 \qquad \forall\, i \neq j.$$
$$V(\epsilon_i) = \sigma^2 \eta(x_i)$$

Slope β of the regression line passing through the origin in the superpopulation model is interpreted as the rate of change Θ in the population. The variance of y is assumed to be an increasing function of x and η is generally a power function: $\eta(x_i) = x_i^{\gamma}$.

According to the model, the best unbiased linear estimator (Brewer 1963 and Royall 1970) of the total is $\hat{Y}_{mc} = \sum_s y_i + \hat{\beta}_{mc} \sum_{\bar{s}} x_i$ where $\hat{\beta}_{mc} = (\sum_s x_i y_i / \eta(x_i)) / (\sum_s x_i^2 / \eta(x_i))^{-1}$ is the least squares estimator.

In the particular case $\eta(x) = x$, this expression reduces to $\hat{\beta}_{mc} = \sum_s y_i / \sum_s x_i$, estimator of the ratio. This unbiased estimator is effective only under the hypothesis of normality of the residuals, and it does not prove to be very robust.

The M-estimators (Huber 1981) serve to define a robust version of the least squares by replacing the square function, in the minimization program, with a function ρ that increases less rapidly:

$$\text{Min} \sum_s \rho\left(\frac{y_i - \beta_R x_i}{\sigma\sqrt{\eta(x_i)}}\right).$$

The M-estimator $\hat{\beta}_R$ is the solution of the following implicit equation:

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

where

$$\psi(t) = \frac{\partial \rho(t)}{\partial t}.$$

The function ψ, like Huber's function $\psi(t) = \text{Max}(-c, \text{Min}(t, c))$, depends on one or more adjustment constants $\underline{c}$ controlling the portion of observations that must be considered as outliers. This estimator will still be sensitive to the effect of outliers on the explanatory variable x. Therefore a more general class of estimators, called GM-estimators (Hampel, Ronchetti, Rousseeuw and Stahel 1986), is defined by means of the following implicit equation:

$$\sum_s w\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right) \psi\left(\frac{r_i}{\sigma} v\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

with

$$r_i = \frac{y_i - \hat{\beta}_R x_i}{\sqrt{\eta(x_i)}}.$$

A choice usually made is Mallows' formulation: $v(t) = 1$ and $w(t) = 1/t$. Hence a robust estimator $\hat{\beta}_R$ will verify the implicit equation

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right) = 0. \tag{4}$$

In general, the parameter σ is unknown and must be replaced in this expression by a robust estimate $\hat{\sigma}$ of the dispersion of the residuals

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = \sum_i \psi\left(\frac{r_i}{\hat{\sigma}}\right) = 0.$$

The estimator of the total will then be:

$$\hat{Y}_{\beta R} = \sum_s y_i + \hat{\beta}_R \sum_{\bar{s}} x_i. \tag{5}$$

This estimator is studied by Gwet and Rivest (1992). In general, it is not unbiased in relation to the sample design. Chambers (1986) proposes to correct that bias by introducing into (5) a third term that estimates it robustly:

$$\hat{Y}_{\text{Chambers}} = \sum_{i \in s} y_i + \hat{\beta}_R \sum_{i \in \bar{s}} x_i +$$

$$\left(\sum_{i \in s} \frac{x_i / \hat{\sigma}\sqrt{\eta(x_i)}}{\sum_{j \in s} x_j^2 / \hat{\sigma}^2 \eta(x_j)} \psi_E\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right)\right) \sum_{i \in \bar{s}} x_i.$$

Choosing a bounded function ψ_E seems a good compromise between estimator bias and variance For example, Welsh and Ronchetti (1994) opt for a Huber's function with a large adjustment constant $c = 15$. But the adjustment of ψ_E, without prior information on the density of the outliers, is always difficult.

3.2 Choice of Estimator

The desirable properties of ψ functions are now well known with reference to the problem of estimating a central tendency. They must be bounded, continuous, and equivalent to an identity in the vicinity of zero. Strictly monotone functions (Huber) are distinguished from redescending functions such as Tukey's biquadratic function, Andrew's sine and the Hampel or Cauchy function. Because their influence function tends toward zero, these estimators will be less sensitive to the presence of outliers than the Huber function. The speed of convergence toward zero is an essential characteristic of redescending functions. Those that are nil at a finite distance (Hampel, Tukey or Andrew) exclude outliers from the estimation of β, whereas the others assign them low representativeness.

The choice and adjustment of the ψ function are difficult. They greatly depend on the nature of the data and more specifically on the distribution of the residuals (Hoaglin, Mosteller and Tukey 1983, Ch. 11). An idea, however approximate, of the appearance of the distribution of the residuals should make it possible to better target both the choice and the adjustment of the estimator, and hence to make the estimation more efficient. This intuitive remark is at the origin of adaptive procedures, presented in particular by Hogg (1974) and (1982). The idea is to evaluate the nature of the distribution of the residuals, calculated on the basis of an initial robust estimate (of the norm L_1 type, for example), using carefully selected robust indicators (tail weight, asymmetry, concentration, *etc.*). The existence of these indicators makes it possible, using a predefined decision rule, to select the appropriate estimator for this situation, and the implicit equation (4) is solved by taking the first robust estimate of β as an initial value.

The idea of an adaptive procedure appears all the more attractive since it systematizes the study that must precede the choice and adjustment of an estimator. That study can prove extremely costly if it must be performed manually for each stratum of the sample and repeated for each survey.

4. CONSTRUCTION OF AN ADAPTIVE PROCEDURE

This section describes the construction of an adaptive procedure for calculating the average rate of change of investment on the basis of economic survey data. Consequently, certain choices were made in light of the specific nature and characteristics of those data and are not necessarily transposable to other regression models. In particular, after checking the data, we adopted the hypothesis of a symmetrical distribution of residuals and we excluded the case of light-tail distributions.

The construction of an adaptive procedure, which draws on the works of Moberg, Ramberg and Randles (1980), is carried out in several stages. The first step is to choose the ψ function (or family of functions) to be used. The second is to select the various criteria for characterizing the distribution of residuals. Using these criteria, a classification rule is constructed. Finally, each class is matched with the adjustment of the estimator to be used.

4.1 Choice of ψ Function

Since Huber-type monotone functions do not provide sufficient protection against outliers, only redescending functions were considered. Among them, we selected the generalized Cauchy function (used in particular by Moberg *et al.* 1980 to approximate generalized lambda functions) and the Tukey biquadratic function:

$$\psi_c(r) = \frac{cr}{(b+r)^2 + c}, \ \forall r$$

and

$$\psi_T(r) = \frac{r}{c}\left(1 - \frac{r^2}{c^2}\right)^2, \ \forall |r| \leq c.$$

These two estimators are quite different in their treatment of outliers (see Figure 1). The biquadratic function equals zero for longer than the Cauchy function, but on the other hand it has a finite rejection point: the residuals beyond $c^*\sigma$ do not enter into the estimate, whereas the Cauchy function assigns them a certain representativeness. The parameter b serves, in principle, to control the asymmetry of ψ according to that of the residuals.

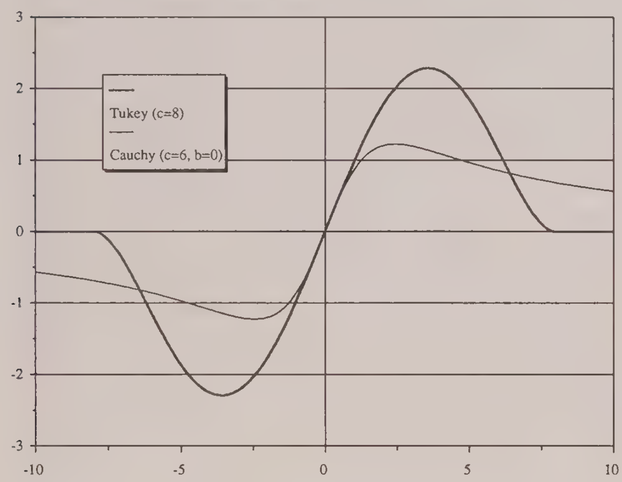


Figure 1. Cauchy and Tukey Functions

4.2 Parameter of Scale, Calculation Algorithm and Selection Criteria

In general an estimator $\hat{\sigma}$ of dispersion is defined by an implicit equation $\sum \chi(r_i/\hat{\sigma}) = 0$, where χ is an even function. It is therefore a matter of solving the system of non-linear equations in $(\hat{\beta}, \hat{\sigma})$ following:

$$\begin{cases} \sum_i \psi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0 \\ \sum_i \chi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0. \end{cases} \qquad (6)$$

Rivest (1989) offers several examples showing that resolving system (6) can pose problems, owing to the fact that there may be a number of solutions, even in the case of a monotone ψ function. Following his recommendations, we will proceed in two stages. First, the parameter of dispersion σ is estimated using the median of the absolute values (MAD) of the residuals defined on the basis of the median of the individual rates of change. Then β is calculated by (4) using the value of σ found previously.

For solving (4), we preferred the reweighting algorithm to the Newton-Raphson algorithm, since it seems to converge more easily, especially when the adjustment constant is small.

Since the effectiveness of an adaptive procedure depends on the effectiveness of the decision-making process, the greatest attention must be paid to the nature, quality and robustness of the information that guides the choice of the estimator. Tail weight is an indispensable indicator, since it provides information on the relative significance of outliers in the sample and thus in the population (see Hoaglin *et al.* 1983, ch. 10). For the tail weight indicator, we adopted the proposal of Hogg (1974):

$$\tau(p) = \frac{\bar{U}(p) - \bar{L}(p)}{\bar{U}(0.5) - \bar{L}(0.5)}$$

$\bar{U}(p)$ (resp. $\bar{L}(p)$) is the mean of the np largest (resp. smallest) order statistics, using a linear interpolation when np is not whole. We chose $p = 0.05$; for the normal distribution $\tau(.05)$ is equal to 2.59.

In addition, like Hogg *et al.* (1988), we considered it important to test for the possible presence of a distribution of the double exponential type, measuring the concentration of residuals by the following pk indicator:

$$pk = \frac{\bar{X}(1-\beta, 1-\alpha) - \bar{X}(\alpha, \beta)}{\bar{X}(.5, 1-\beta) - \bar{X}(\beta, .5)}$$

where $\bar{X}(a, b)$ is the means of the order statistics between the na-th and the nb-th, with the sizes interpolated if na or nb are not integers. We selected $\alpha = 0.05$ and $\beta = 0.15$, or $pk = 2.7$ for a normal distribution.

Finally, different studies (Moberg *et al.* 1980, Hogg *et al.* 1988) have emphasized the importance of the dissymmetry of distributions. When there are asymmetrical residuals, the bias of robust estimators can be sizable, making it tricky to use them (Chambers et Kokic 1993). In the INSEE Investment Survey, the residuals are theoretically asymmetrical since they are confined to a limited range ($r = y - \beta x \geq -\beta x$). However, we noted empirically that this asymmetry was very slight and could safely be ignored. The failure of the correction of a possible bias by the function ψ_E in Chambers' estimator moreover confirms this observation. Only the symmetrical case is considered here; the bias of the estimators defined by (5) is therefore nil.

4.3 Classification of Distributions and Adjustment of the Estimator

The definition of the decision rule was based on the study of eight specific symmetrical distributions illustrating various tail weight and concentration situations (see Table 1). We were interested in the family of contaminated distributions $CN(\alpha, K)$, with the distribution function $F(x) = (1-\alpha)\Phi(x) + \alpha\Phi(x/K)$ where Φ is the cumulative function of the distribution $N(0, 1)$, since these distributions give a good representation of real data (Hoaglin *et al.* 1983, ch. 10), especially the data in the Investment Survey (Ravalet 1996). While Gaussian in the middle, they nevertheless contain more outliers than the normal distribution $N(0, 1)$.

Table 1
Eight Specific Distributions

		$\tau(.05)$	pk
1	Normal distribution	2.59	2.76
2	Contaminated dist $CN(.05, 3)$	2.94	2.83
3	Double exponential dist.	3.28	3.41
4	Contaminated dist $CN(.05, 10)$	4.47	2.85
5	Contaminated dist $CN(.10, 10)$	5.42	3.05
6	Contaminated dist $CN(.20, 10)$	5.64	4.44
7	Slash distribution	7.65	4.19
8	Cauchy distribution	7.82	4.78

The two indicators $\tau(0.5)$ and pk were simulated over these eight distributions, for several sample sizes. The graph of $(\tau(0.5), pk)$ serves to distinguish four groups of distributions: light-tailed, relatively unconcentrated distributions of the normal type or $CN(.05,3)$; heavy-tailed distributions of the type $CN(.05,10)$, $CN(.10,10)$, and $CN(.20,10)$, and very heavy-tailed distributions of the Slash or Cauchy type; and concentrated distributions such as the double exponential distribution. These four classes are defined (see Figure 2) by the following equation boundaries:

$$\text{Class I:}\quad \tau(0.5) \le 3.6 - \frac{14}{n} \quad \text{and} \quad pk \le 3.20$$

$$\text{Class II:}\quad 3.6 - \frac{14}{n} < \tau(0.5) \le 5.8 - \frac{35}{n}$$

$$\text{Class III:}\quad 5.8 - \frac{35}{n} < \tau(0.5)$$

$$\text{Class IV:}\quad \tau(0.5) \le 3.6 - \frac{14}{n} \quad \text{and} \quad pk > 3.20$$

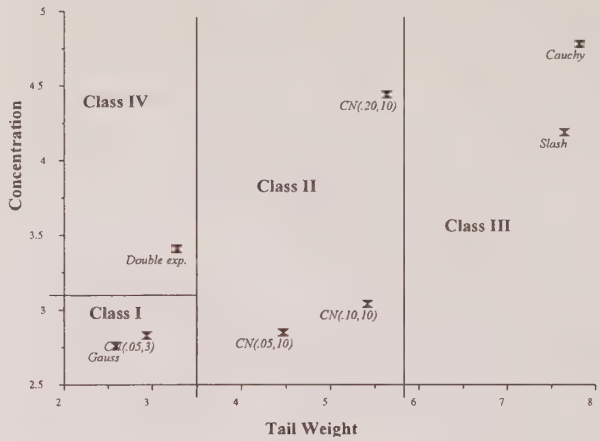


Figure 2. Four Classes of Distributions

The final stage consists in setting the adjustment of the two estimators in each class. Since we are interested only in the symmetrical case, the b parameter of the Cauchy function is nil. By simulations, we determined for the eight reference distributions the optimal constants c of the Tukey and Cauchy functions (*i.e.*, minimizing the variance of these estimators or, what amounts to the same thing here, their mean square deviation). These do indeed diminish with tail weight, except of course for the case of the double exponential distribution, which requires an adjustment similar to those used for the Slash and Cauchy distributions.

Tukey's estimator is more efficient on the normal or contaminated distributions, but it generally requires finer adjustment. Figure 3 shows the example of the contaminated distribution *CN(.10,10)*. Lastly, while the choice of the constant appears to be relatively critical for the heavy-tailed or concentrated distributions, a wide band of value is possible for distributions close to the normal distribution.

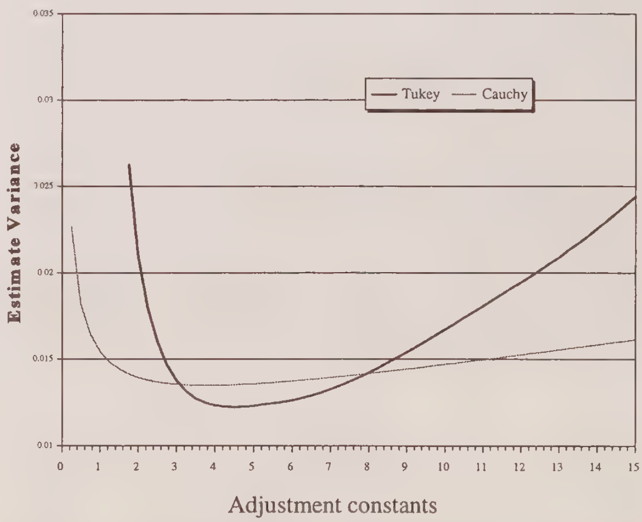


Figure 3. Variance of Tukey and Cauchy Estimators for the Distribution CN(.10,10) (n=100)

The synthesis of these results serves to define the adjustments to be used on each distribution class. These adjustments, established for samples of size 100 (Table 2), remain entirely acceptable for samples sizes between 50 and 150.

Table 2
Adjustment of Estimators by Class of Distribution of Residuals (n = 100)

Class	Tukey	Cauchy
I	7	7
II	4.5	4
III	3	1
IV	3	1

5. APPLICATION TO THE INVESTMENT SURVEY

5.1 The Problem of Stratification

The strata used for the LI estimator are defined by the cross-tabulation of an activity (18 manufacturing sectors) and a company size class (small, medium or large). Among these 54 strata, approximately 20 never contain more than 20 observations. This stratification is therefore too fine for the adaptive procedure to be used correctly, as it assumes a minimum number of observations.

Since small firms are fairly distinct from medium-sized and large firms in terms of dispersion and residuals tail weight, differentiation by size is maintained. Sectors must thus be grouped. We decided not to adopt the method used by Sohre (1995), which consists of grouping after data collection those sectors having the closest parameters (here the average change in investment). Proximity is impossible to assess in small strata, and the groups obtained are likely to change from one survey to another, making comparisons difficult. We preferred to redefine 15 new strata based on a higher classification level distinguishing only four sectors: intermediate goods, professional capital goods, automobile, and consumer goods.

5.2 Characteristics of Strata

The hypothesis of a variance of residuals independent of x in the model ξ cannot be accepted. The choice of γ in the function η is made in such a way that the curve of the residuals (in absolute value) as a function of the regressor, smoothed by the LOESS method, shows no trend (Cleveland 1979). For the stratum representing intermediate goods and medium-sized companies in the April 1995 survey (see Figure 4), $\gamma = 1.3$ is an acceptable compromise between the appearance of a downward trend for small values of x and the cancellation of the upward trend for the larger values of x. A similar examination on the other strata confirmed this choice for the manufacturing industry as a whole.

In each stratum, the distribution of the residuals systematically exhibits a heavier tail than the normal distribution, without being extremely heavy-tailed. Within a given sector, the tail weight indicator decreases with company size. The great majority of the strata representing small and medium-sized firms were assigned to Class 2. Large firms more often exhibit somewhat heavy-tailed distributions, close either to the normal distribution (Class 1), or the double exponential distribution (Class 4). Class 2 is by far the largest and represents 75% of cases. Only 20% of the distributions are recognized as somewhat heavy-tailed and are assigned in equal proportions to classes 1 and 4. On the other hand, very heavy-tailed distributions (Class 3) are unusual (less than 5% of the cases). While there appears to be a certain persistence to the classification, it is not perfect. And the changes are quite real, since they resist a slight modification of the boundaries between classes. Thus this perfectly justifies the use of an adaptive procedure.

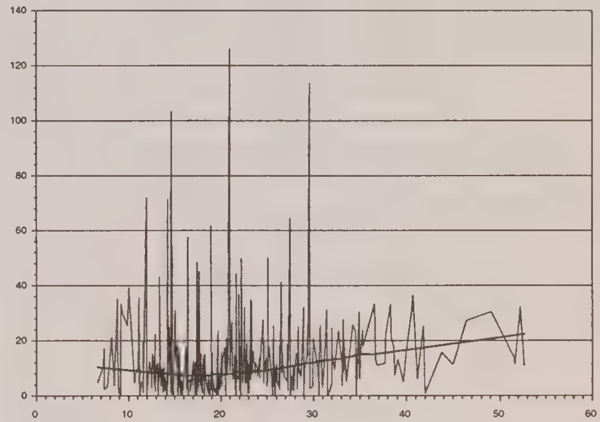


Figure 4. Absolute Value of Residuals ($\gamma = 1.3$, Intermediate Goods, Size 2, April 95)

5.3 Resulting Estimates

The estimation procedure based on (5), applied to the six surveys covering the period 1990-1995, yielded the results shown in Figure 5. Also shown are National Accounts estimates, those obtained with the LI estimator, and those from the Annual Business Survey (ABS), which is exhaustive.

For the manufacturing sector as a whole, the results of the adaptive procedure are comparable to those obtained with the LI estimator. The biquadratic function results in estimates that are consistently lower than those obtained with the Cauchy function. With a finite rejection point, the Tukey function is less influenced by the slight asymmetry toward the right in the distribution of the residuals. These new estimates are closer to those of the ABS than to the National Accounts estimates. This is hardly surprising, considering the excellent correlation between individual

ABS data and the responses obtained in the survey. As yet there is no explanation for the differences in 1991 and 1994 in relation to the National Accounts estimates. Apart from the year 1994, the estimates obtained with the Cauchy function are entirely acceptable in the intermediate goods and automobile sectors and to a lesser extent in the professional capital goods sector. On the other hand, in consumer goods, the results are fairly far from the National Accounts estimates. Here we are likely running up against a problem of sample quality. This sector is quite heterogeneous, and a few activities such as printing are poorly covered by the survey.

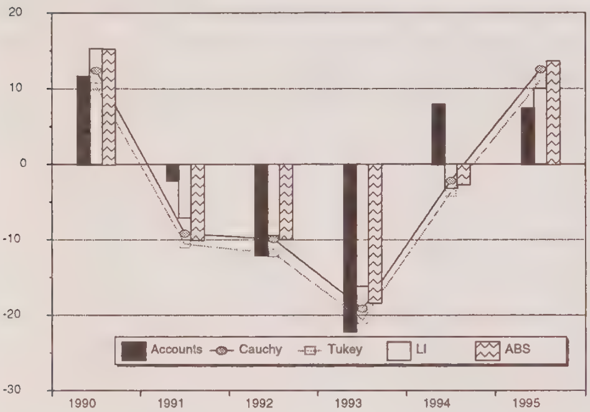


Figure 5. Investment Growth Rate in Value in the Manufacturing Industry

6. CONCLUSIONS

This article presents a theoretical justification of a procedure currently used to process data from the Investment Survey; in particular it offers a justification of the principle of excluding outliers or large investors. However, the strategy of reweighting the linear estimator following Hidiroglou and Srinath (1981) shows itself to be insufficient for this purpose in several respects, mainly having to do with the identification and treatment of representative outliers. The dichotomy between extrapolatable individuals and large investors appears too radical and leads to a lack of robustness, since the influence curve of this estimator is not continuous.

On the other hand, the hypothesis of a linear superpopulation model and its estimation by GM-estimators seemed to us to be of great interest from both a methodological and practical standpoint. The insertion of these techniques into an adaptive procedure also makes it possible to have a robust estimator for a variety of situations. Following principles described in the literature, the procedure proposed here uses indicators of tail weight and concentration of the residuals in the linear model calculated from the sample, to decide on the adjustment of the weight function to be used, it being assumed that the residuals are symmetrical. The estimates made with the Cauchy function yielded satisfactory results on the manufacturing industry, and they largely validate previously published results. The advantages of this method over the one currently used basically have to do with lower implementation costs and greater control over the methodology employed.

The adaptive procedure was constructed independently of the survey, and therefore there is no guarantee that the classification is optimal for the strata content. Furthermore, we did not study the robustness of the rule for assigning values to a class. This issue is important when one carries out several successive measurements and one wants to interpret the revisions. Clearly, further research on these classification methods is required, in order to integrate additional information such as the information yielded by earlier estimates or comprehensive surveys of the population studied.

ACKNOWLEDGEMENTS

The author wishes to thank Michel Hidiroglou and Dominique Ladiray for their comments and suggestions during the preparation of this article.

REFERENCES

BREWER, K.R. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of Statistics*, 5, 93-105.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

CHAMBERS, R.L., and KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute, Proceedings of the 49th Session, Book 2*, 55-72.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GWET, J.P., and RIVEST, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E., ROUSSEEUW, P.J., and STAHEL, W.E. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: John Wiley.

HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

HOGG, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of American Statistical Association*, 69, 909-923.

HOGG, R.V. (1982). On adaptive statistical inferences. *Communication in Statistics*, 11, 2531-2542.

HOGG, R.V., BRIL, G.K., HAN, S.M., and YUL, L. (1988). An argument for adaptive robust estimation. *Probability and Statistics Essays in Honor of Franklin A. Graybill.* Amsterdam: North-Holland/Elsevier, 135-148.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1995). Outliers in business surveys. In *Business Survey Methods*. New York: John Wiley.

MOBERG, T.F., RAMBERG, J.S., and RANDLES, R.H. (1980). An adaptive multiple regression procedure based on M-estimators. *Technometrics*, 22, 213-224.

RAVALET, P. (1996). L'estimation du taux d'évolution de l'investissement dans l'enquête de conjoncture: analyse et voie d'amélioration. Document de travail de l'INSEE Méthodologie Statistique, 9604.

RIVEST, L.P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de régression sont estimés simultanément. *Canadian Journal of Statistics*, 17, 141-153.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SOHRE, P. (1995). The Adaptive KOF Procedure for the Estimation of Industry Investment. 22nd CIRET Conference, Singapore.

WELSH, A.H., and RONCHETTI, E. (1994). Bias-Calibrated Estimations of Totals and Quantiles From Sample Surveys Containing Outliers. Technical Report, Dept. of Econometrics, University of Geneva, Switzerland.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 109-117
Statistics Canada

Sampling and Maintenance of a Stratified Panel of Fixed Size

F. COTTON and C. HESSE[1]

ABSTRACT

Statistical agencies often constitute their business panels by Poisson sampling, or by stratified sampling of fixed size and uniform probabilities in each stratum. This sampling corresponds to algorithms which use permanent numbers following a uniform distribution. Since the characteristics of the units change over time, it is necessary to periodically conduct resamplings while endeavouring to conserve the maximum number of units. The solution by Poisson sampling is the simplest and provides the maximum theoretical coverage, but with the disadvantage of a random sample size. On the other hand, in the case of stratified sampling of fixed size, the changes in strata cause difficulties precisely because of these fixed size constraints. An initial difficulty is that the finer the stratification, the more the coverage is decreased. Indeed, this is likely to occur if births constitute separate strata. We show how this effect can be corrected by rendering the numbers equidistant before resampling. The disadvantage, a fairly minor one, is that in each stratum the sampling is no longer a simple random sampling, which makes the estimation of the variance less rigorous. Another difficulty is reconciling the resampling with an eventual rotation of the units in the sample. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used for the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the numbers following a uniform distribution.

KEY WORDS: Panel; Stratified sampling of fixed size; Stratified simple random sampling; Maximum coverage; Sample rotation; Equidistant numbers.

1. INTRODUCTION

We consider the successive selection of samples intended to follow the change over time of sums of variables, more generally functions of sums, in a population. For example, this may be a population of businesses or establishments for which we wish to follow monthly sales trends. The ideal would be to be able to conserve a constant sample, but demographic movements make this impossible and it may not be desirable in light of the survey response burden.

The methods for selecting units presented in this article are subject to the following three constraints:

Firstly, it is necessary to regularly introduce births and to take deaths into account.

Secondly, sampling involves characteristics of units which change over time, such as the size or primary activity of businesses. These characteristics can be used to modulate the probabilities of inclusion. Notably, it is often prudent to increase these probabilities with the size of the units if we estimate sums of variables correlated with this size. In addition, these characteristics may eventually be used as stratification criteria. In this article, a stratum will mean a subset of the population within which the sampling is of fixed size, to the nearest rounded digit. However, the criteria used in the stratification of the first sampling, such as the primary activity of the unit, become "inexact" or become less and less correlated with the variables of interest such as size. This results in a progressive increase in the variance of the estimates. To remedy this, it is appropriate to carry out a resampling of the sample from time to time after updating the stratification and calculating new probabilities of inclusion. This must be done while endeavouring to conserve the maximum number of units. However, fatally, units will be excluded and others will be introduced, mainly because of changes in the probabilities of inclusion, although this would also happen because of the changes of strata, even if the probabilities of inclusion remained constant.

Thirdly, we would like to distribute our survey response burden over a larger number of units. We determined a maximum duration limit for inclusion in the panel. Beyond this limit, the unit is replaced by another unit chosen from those which have never been included, or which have been absent the longest. We call this change of the sample over time rotation. It is generally slow and regular. The various methods for performing this rotation are well known in statistical agencies. They consist mainly in attributing, at the beginning, a permanent random number to each unit of the population. The successive samples are defined by intervals over these numbers or by the ranks induced by these numbers.

We call the chronological sequence of samples resulting from these updating operations a "panel" and the set of updating operations "maintenance" of the panel.

The maintenance scheme presented in this article is analogous to that of Hidiroglou, Choudhry and Lavallée (1991). It corresponds to a frequency of updating of the

[1] F. Cotton, Institut National de la Statistique et des Études Économiques, Département de l'Informatique and C. Hesse, Institut National de la Statistique et des Études Économiques, Département "Système Statistique d'Entreprises", 18 boulevard Adolphe-Pinard, 75675, Paris Cedex 14.

stratification and probabilities which is significantly less than the survey frequency. This is generally the case for surveys with an infra-annual periodicity. The speed of demographic movements is not considered large enough to make it worthwhile to reselect the sample every time. The rotation is carried out without changing the probabilities of inclusion and the strata between two resamplings and it is regularly spread over time to conserve a certain continuity of the quality of the estimators of change over time. This also corresponds to a duration of inclusion of which the expected value is constant. In certain algorithms, we could determine a constant duration between two resamplings; otherwise we could set an upper limit. The speed of rotation represents a compromise between the efficiency of the estimators of change over time, which is greater the lower the rate of renewal, and the concern not to keep a unit in the panel for too long. Note that the quest for maximum coverage in the resampling remains meaningful with the rotation: we first remove the fraction to be renewed as if there were no resampling, then we seek the maximum coverage with the residual portion.

We will examine several methods of panel maintenance, with emphasis on maximizing sample coverage during resamplings. We will distinguish more particularly a process which assigns equidistant numbers to the units before each change of stratum.

The article is divided as follows:

After reviewing definitions and describing a few notations in section 2, we briefly indicate in section 3 how Poisson sampling makes it possible to carry out the previous maintenance scheme simply and perfectly. This sampling has the disadvantage of being of random size, but it serves as a reference for the stratified sampling of fixed size which we then consider.

In most instances, in these samplings, we determined probabilities of inclusion at the outset and used a rounded number to determine an entire sample size in each stratum. This problem, examined in section 4, is not negligible when the strata are small, which can occur for strata of births. In addition, rounding is used in the method which we propose to maximize the coverage after resampling.

Section 5 deals with the maximum coverage of samples of fixed size. First, we review two known methods: that of Kish and Scott (1971) and another based on the attribution to each unit of permanent independent numbers following the uniform distribution. The Kish and Scott method (1971) seems poorly suited to an intermediate rotation between resamplings. The other method, which reproduces simple random sampling in each stratum, does not have this disadvantage, but the coverage is less than with the Kish and Scott method (1971). Finally, we propose that the numbers be equidistant before resampling. We then obtain the same coverage as with the Kish and Scott method (1971), at least in the case of proportional distribution, while facilitating intermediate rotations. However, the coverage remains less than the maximum theoretical coverage which we obtain, for example, with Poisson sampling.

In sections 6 and 7, we present the intermediate phases of updating births and deaths and of rotation.

To conclude the topic of maintenance, we show in section 8 how resampling can take place between two phases of rotation. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used in the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the uniform beginning numbers if we wish to continue with simple random sampling.

2. REMINDERS, DEFINITIONS AND NOTATIONS

Let there be a population, or finite set of units $i \in U = \{1, ..., N\}$ where N is the size of the population.

We consider only samples without replacement. A sample is then simply a subset s of U. We call sample size the number n of units which it contains.

A sampling or selection plan is a discrete probability $p(s)$ over the set of samples.

We can generalize to joint sampling of several samples. By limiting ourselves to two samples s_1, s_2, the joint sampling is the probability $p(s_1, s_2)$ over the set of pairs (s_1, s_2).

The first-order probability of inclusion of an individual i is defined by:

$$\pi_i = \sum_{s \ni i} p(s).$$

$E(.)$ being the expected value with respect to the sampling, this yields:

$$E(n) = \sum_{i \in U} \pi_i .$$

In the case of two samples with first-order probabilities of inclusion $\pi_{i,1}, \pi_{i,2}$, we can define the joint probability of inclusion:

$$\pi_{i,1,2} = \sum_{s_1 \ni i,\, s_2 \ni i} p(s_1, s_2).$$

This yields the constraint:

$$\pi_{i,1,2} \le \min(\pi_{i,1}, \pi_{i,2}). \tag{2.1}$$

If $i \in s_1$, the probability of reselection in s_2 is $\pi_{i,1,2}/\pi_{i,1} \le \min(1, \pi_{i,2}/\pi_{i,1})$.

In Poisson sampling, the selection of the units is independent and the sample size is random. Except in section 3, we will instead consider sampling where the size is fixed to the nearest rounded digit.

Simple random sampling (SRS) is sampling of fixed size where the samples are equiprobable. This yields $\pi_i = n/N$.

The population is partitioned into strata $U_h, h = 1, ..., H$ of sizes N_h. In this article, we will call a set of H independent samples of fixed size n_h in each stratum

"stratified sampling of fixed size" and we will limit ourselves to samplings with a uniform first-order probability of inclusion in each stratum. We will then use the notation $f_h = \pi_i$. We will call a stratified sampling of fixed size with simple random sampling in each stratum "stratified simple random sampling" (SSRS).

We will call the number of consecutive surveys where a unit is included in the panel "duration of inclusion of a unit." We will notate it D_i, or D_h in the particular case where it is the same for all units of a stratum h. When $\pi_i \geq 0.5$, this duration cannot be less than $\pi_i/(1-\pi_i)$. For example, if $\pi_i = 0.7$, the duration of inclusion is at least 3. In practice, we will not rotate units whose π_i exceeds a certain threshold.

In addition, the previous variables are indexed by survey wave t. The population U_t of size N_t and the sample s_t of size n_t vary because of births and deaths, and the sample also varies as a result of the stipulated rotation. Moreover, we will consider samples at particular times $t = t_1$ of the first sampling and $t = t_2$ of the first resampling. For the sake of simplicity, they will be notated s_1, s_2 instead of s_{t_1}, s_{t_2}. The algorithms described for the pair (s_1, s_2) will be valid for the following resampling pairs.

3. SOLUTION BY POISSON SAMPLING

It is enlightening to examine how we can observe the panel maintenance scheme by Poisson sampling. This is the model which we will endeavour to approximate in order to choose a selection method.

We attribute to each unit i, at its birth, a number which is a random number ω_i selected according to the uniform distribution in [0,1). It is implicit in the formulae where these numbers appear that the results of the operations are *modulo* 1.

During the first sampling, at date $t = t_1$, we select the units such that ω_i belongs to the interval $[0, \pi_{i,1})$ where $\pi_{i,1}$ are the probabilities of inclusion given. In the absence of rotation, we keep this interval at the following dates until resampling. Births as well as deaths are distributed at random in this interval. The resampling, at date $t = t_2$ is carried out by selecting the units of the interval $[0, \pi_{i,2})$ where $\pi_{i,2}$ are new probabilities of inclusion. The joint probability of inclusion is equal to the length of the common interval, *i.e.*, $\min(\pi_{i,1}, \pi_{i,2})$ which is the maximum theoretically possible according to the formula (2.1). The expected value of the coverage is therefore itself maximal.

Let us now consider a rotation between the sampling and the resampling. We maintain the probability $\pi_{i,1}$ and we can determine a duration of inclusion $D_{i,1}$, which is variable depending on the units, but fixed until the resampling. This constraint is realized by defining the sample at date $t(t_1 < t < t_2)$ by the interval

$$\omega_i \in [(t-t_1)\pi_{i,1}/D_{i,1}, (t-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}).$$

The rate of rotation is a random variable. Its expected value results from $D_{i,1}$. It is equal, for any subset V of the population, to $\sum_{i \in V}(\pi_{i,1}/D_{i,1})/\sum_{i \in V}\pi_{i,1}$.

At the first resampling at date $t = t_2$, we could define the sample by

$$\omega_i \in [(t_2-t_1)\pi_{i,1}/D_{i,1}, (t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}).$$

However, we encounter a difficulty for units such that

$$\pi_{i,2} < \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right),$$

and if ω_i belongs to the interval

$$\left[(t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}, (t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right)\right).$$

These units, which were previously in the sample, are excluded but will be reincluded in a future rotation. If we wish to avoid this, we must make the limit of the new interval coincide with that of the old interval, and the sample at date $t = t_2$ is finally defined by:

$$\omega_i \in [a_{i,1}, a_{i,1} + \pi_{i,2}),$$

where:

$$a_{i,1} = (t_2-t_1)\pi_{i,1}/D_{i,1} + \max\left[0, \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right) - \pi_{i,2}\right].$$

The joint probability of inclusion is equal to the length of the common interval, *i.e.*,

$$\min\left(\pi_{i1}\left(1 - \frac{1}{D_{i,1}}\right), \pi_{i,2}\right).$$

This is also the maximum compatible with the rotation.

If we continue the rotation with durations of inclusion $D_{i,2}$ the interval at date $t > t_2$ is:

$$\left[a_{i,1} + (t-t_2)\pi_{i,2}/D_{i,2}, a_{i,1} + (t-t_2)\pi_{i,2}/D_{i,2} + \pi_{i,2}\right).$$

Poisson sampling controls exactly the duration of inclusion and maximizes, as an expected value, the coverage during resampling but with the disadvantage of a random sample size, regardless of the subpopulation. In the following pages, we will endeavour to devise algorithms similar to those just described for Poisson sampling in order to apply them to stratified sampling of fixed size. We will try to control the duration of inclusion in the rotation, as for Poisson sampling, and to approximate the same rate of coverage during resampling. We will begin with the problem of coverage during resampling in section 5, but first, it is useful to clarify certain concepts concerning the rounding of sample sizes by stratum.

4. ROUNDING OF SAMPLE SIZES BY STRATUM

This problem is related to the estimation formulae. These formulae use the first-order probabilities of inclusion, either in the unbiased Horvitz-Thompson estimator or in adjusted estimators. Let f_h be the probability of inclusion by stratum, and let $v_h = N_h f_h$. We must have a whole number n_h per stratum. An initial method for accomplishing this consists in restricting the choice of the f_h in such a way that v_h is an integer. In each stratum where we would have had $v_h < 1$, we must take $v_h = 1$ so that $f_h > 0$. However, if the stratification is very fine vis-à-vis the sample size, this occurs in numerous strata. This makes it necessary either to increase the sample size or to decrease the sampling rate in the other strata, to the detriment of efficiency.

We will use a second method, which consists in linking the probability f_h more loosely to n_h. We apply a rounding process such that $E(n_h) = v_h$, where v_h is no longer necessarily an integer.

Let us assume that $I(.)$ is the integer part function. We must have

$$\Pr[n_h = I(v_h) + 1] = \varphi_h,$$

$$\Pr[n_h = I(v_h)] = 1 - \varphi_h,$$

where $\varphi_h = v_h - I(v_h)$.

It is then no longer necessary that $n_h > 0$ in order for $f_h > 0$. Note that the first method can be considered a particular case of the second. This rounding can be done independently by stratum, in a linked way by systematic rounding or by the Cox method (1987). We describe only systematic rounding.

Let us first order all of the strata, and index them by their rank. Let $c_0 = 0$ and $c_h = \sum_{j=1}^{h} \varphi_j$; we select a number θ in the interval [0, 1), according to the uniform distribution and we take $n_h = I(v_h) + 1$ in the strata such that $c_{h-1} \leq m - 1 + \theta < c_h$ for m entirely.

This implies that

$$|(n_{j_1} + \ldots + n_{j_2}) - (v_{j_1} + \ldots + v_{j_2})| < 1,$$

for any j_1, j_2 such as $1 \leq j_1 \leq j_2 \leq H$.

In particular, the global size differs by less than one unit from its expected value. This is obviously not the case with independent roundings.

5. ALGORITHMS FOR THE MAXIMUM COVERAGE OF SAMPLES OF FIXED SIZE

The maintenance algorithms which we propose are based on the attribution of equidistant numbers. This is not necessary during the first sampling, nor in the rotation, but is used to maximize the coverage during updates of the stratification. That is why we examine this maintenance phase first.

Let us begin by describing all the notations and making a few useful observations.

We select a first sample s_1 stratified according to criterion h_1. After a certain time has elapsed, we select a new sample s_2 with an updated stratification h_2. The first-order probabilities of inclusion are respectively f_{h_1}, f_{h_2} and the sample sizes required by stratum are respectively n_{h_1}, n_{h_2}. It is sufficient to consider what happens in any stratum $h_2 = g$. Let $s_{g,1}$ be the part of the first sample s_1 in this new stratum, of which the size $n_{g,1}$ is generally random. Let $s_{g,2}$ be the part of the second sample s_2 in this new stratum, of which the size is fixed to the nearest rounded digit. The size $n_{g,1,2}$ of the coverage cannot exceed the limit $\overset{+}{n}_{g,1,2} = \min(n_{g,1}, n_{g,2})$. We can hope to devise $s_{g,2}$ a resampling process with a uniform first-order probability of inclusion in $s_{g,1}$ which makes it possible to attain this limit, at least when the first-order probabilities of inclusion in are also equal to a single value $f_{h_1} = f_1$. Note that, even if this limit is attained, the fixed size constraints decrease the coverage. The finer the stratification, the greater this effect. In fact, the smaller the population of stratum g, the greater the likelihood that the coefficient of variation of $n_{g,1}$ will be large, as well as the proportion of units not reselected in the case $n_{g,1} > n_{g,2}$.

There is an obvious way of attaining the limit $\overset{+}{n}_{g,1,2}$. Let us assume first of all that the first-order probabilities of inclusion in $s_{g,1}$ are uniform. If $n_{g,1} < n_{g,2}$, we add $n_{g,2} - n_{g,1}$ units to $s_{g,1}$ selected at random in the complement of $s_{g,1}$. If $n_{g,1} > n_{g,2}$, we remove $n_{g,2} - n_{g,1}$ units from $s_{g,1}$ selected at random. By construction this yields $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$, and $n_{g,1,2} = \overset{+}{n}_{g,1,2}$. If the first-order probabilities of inclusion in $s_{1,g}$ are not uniform, we apply the same method within subsets where these probabilities are uniform. This is the method proposed by Kish and Scott (1971) on page 468 of their article. They do not stipulate the procedure for random selection, but we assume that it is SRS.

As Kish and Scott point out, the second-order probabilities of inclusion are not uniform and if the first sampling is a SSRS, the second sampling no longer meets this definition. The first-order probability of inclusion, itself, is not strictly uniform when includes elements of strata from the previous sampling: see an example in the appendix. However, there is another method which verifies this condition. It is well known to statistical agencies which practise coordination of samples. For the sake of convenience, we will call it "method 1".

Method 1: Use of independent numbers following the uniform distribution

We attribute to the units, at their birth, ω_i numbers which follow the uniform distribution in [0, 1) and are independent, as in Poisson sampling. The first sample s_1 is obtained by selecting, for example, the n_{h_1} units of lower rank according to ω_i in each stratum. With this algorithm, the maximum coverage is also obtained by selecting the n_{h_2}

units of lower rank according to ω_i in each stratum h_2. Moreover, it is obvious that these two samplings are SSRS.

It is also obvious that we cannot obtain greater coverage with this algorithm. In addition, we conjecture that it is not possible to do better, for SSRS, regardless of the algorithm.

On the other hand, the coverage is poorer as an expected value than with the Kish and Scott method (1971), at least in the particular case where the first-order probabilities of inclusion in s_1 are uniform. In fact, at that point the relations $g\, s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}, n_{g,1,2} = n^+_{g,1,2}$, are not necessarily true and the loss of coverage is greater, the smaller the strata during the first sampling.

We shall demonstrate this, again in the particular case of a uniform probability of inclusion f_1 in s_1. Let us assume that ω_{h_1} is the greatest value of ω_i for the units of s_1 in stratum h_1, and ω_g the greatest value of ω_i for the units of s_2 in stratum g. Let $\omega_1^- = \min(\omega_{h_1})$ and $\omega_1^+ = \max(\omega_{h_1})$. If $\omega_g \le \omega_1^-$ then $s_{g,2} \subseteq s_{g,1}$ and if $\omega_g \ge \omega_1^+$, then $s_{g,2} \supseteq s_{g,1}$. In both cases $n_{g,1,2} = n^+_{g,1,2}$. The risk of not attaining the limit exists only if $\omega_1^- \le \omega_g \le \omega_1^+$. In this case, the relation $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$ is no longer necessarily true: see Figure 1, where we considered only 2 strata h_1. The loss of coverage is greater where the quantity $\omega_1^+ - \omega_1^-$ is greater as an expected value, and therefore where the strata h_1 are smaller.

Method 2:
Use of equidistant numbers

If we accept not to conserve a SSRS, how can we modify the previous method to obtain the same coverage as the Kish and Scott method (1971), at least when we have the uniform probability of inclusion f_1 in s_1? We have seen that the loss of coverage was the result of the deviation between the ω_{h_1}. It is sufficient to transform the ω_i into new numbers $\rho_{i,1}$ in such a way that the ρ_{h_1} which correspond to the ω_{h_1} are as close as possible to a common value, *i.e.*, f_1. More specifically, we would like to have the equivalence:

$$\{i \in s_1 \Leftrightarrow R_{h_1}(i) \in [1, ..., n_{h_1}]\} \Leftrightarrow \rho_{i,1} \in [0, f_{h_1}),$$

where $R_{h_1}(i)$ is the rank according to ω_i in h_1 of unit i. A solution is given by the transformation:

$$\rho_{i,1} = \frac{R_{h_1}(i) - 1 + \theta_{h_1}}{N_{h_1}} \tag{5.1}$$

where θ_{h_1} is a real number which verifies:

$$\begin{cases} \theta_{h_1} \in [0, \varphi_{h_1}), n_{h_1} = I(v_{h_1}) + 1, \\ \\ \theta_{h_1} \in [\varphi_{h_1}, 1), n_{h_1} = I(v_{h_1}). \end{cases}$$

The transformation therefore involves the rounded number of the v_{h_1} examined in section 4. The sampling of s_2 is carried out like that of s_1 except that the $\rho_{i,1}$ now play the role of the ω_i: in each new stratum g we define rounded sizes $n_{g,2}$ and we select the $n_{g,2}$ units of lower rank according to $\rho_{i,1}$. Note that these ranks are different from those induced by ω_i.

Let us assume that the probability of inclusion in s_1 is still uniform. Let ρ_g be the value of $\rho_{i,1}$ for the unit of rank $n_{g,2}$ in g. If $\rho_g \in [0, f_1)$, then $s_{g,2} \subseteq s_{g,1}$. Otherwise $s_{g,2} \supseteq s_{g,1}$. In this particular case, we therefore attain the maximum coverage $n^+_{g,1,2}$ as in the Kish and Scott method (1971), and unlike method 1. We illustrate in Figures 1 and 2 how the transformation into equidistant numbers makes it possible to increase the coverage compared to method 1.

We apply the same algorithm when the probabilities of inclusion in s_1 are not uniform. Unlike the Kish and Scott method (1971), we do not need to fix the size of the new sample within subsets where these probabilities are uniform. This is another advantage and we think that it increases the coverage.

Nonetheless, the coverage obtained by this algorithm remains lower, as an expected value, than that of a Poisson sampling with the same probabilities of inclusion. In order to have, as an expected value, the same coverage as with Poisson sampling, it would be sufficient to define $s_{g,2}$ by $\rho_{i,1} \in [0, f_g)$. In fact, we would then have $\Pr(i \in s_i \cap s_2) = \min(f_{h_1}, f_g)$, but the sampling so obtained would no longer be of fixed size.

The following resamplings, after new updates, are carried out by repeating the process. For example, before selecting s_3 we calculate equidistant numbers $\rho_{i,2}$ based on $\rho_{i,1}$ (and not ω_i) in each stratum h_2.

The resulting sampling plan in the new strata is no longer a SRS. In particular, the probabilities of inclusion of the pairs of units vary generally as a function of the former strata. In other words, the resampling keeps a "trace" of the stratification of the first sampling. Moreover, the probabilities of inclusion of the units in $s_{g,2}$ are not exactly equivalent to f_g, except for the sample defined by $\rho_{i,1} \in [0, f_g)$. For the sample of fixed size $n_{g,2}$ this probability varies as a function of the size of the former strata. As in the Kish and Scott method (1971), we do not strictly control these probabilities. However, the deviation between f_g and the true probability becomes negligible when $n_{g,2}$ is sufficiently large.

Note 1. The transformation of numbers which independently follow the uniform distribution in equidistant numbers was proposed by Brewer, Early and Hanif (1984) as a way of rotating samples in the same manner as Poisson sampling, with the advantage of a smaller variance of the sample size. However, this transformation is performed by taking the set of the population, and therefore they did not address the problem of maximum coverage during changes of stratum. The numbers change only when births and deaths are updated, according to a procedure which is also quite different from that which we propose for changes of stratum.

Note 2. In the demonstration we just provided, it is not necessary that the numbers be completely equidistant. It is sufficient that the n_{h_1} units of s_1 and the $N_{h_1} - n_{h_1}$ complementary units have their new numbers respectively in $[0, f_{h_1})$, $[f_{h_1}, 1)$. We could attribute these new numbers

in such a way that they independently follow the uniform distribution in these intervals.

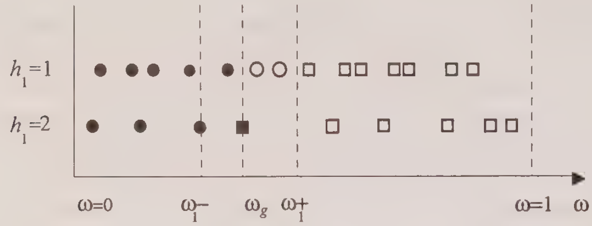


Figure 1. Coverage with method 1 (numbers following the uniform distribution).

We have represented the units in g according to the value of the number ω (on the abscissa) and the stratum h_1 of the first sampling (on the ordinate). We assume that there are only two strata. The circles correspond to $s_{g,1}$ and the squares to the complementary part. The solids correspond to $s_{g,2}$ and the blanks to the complementary part. The size of $s_{g,2}$ was fixed at 9 which defines ω_g. In this example, we see that two units are not reselected (in $h_1 = 1$) and that another is new (in $h_1 = 2$). The size of the coverage is 8, while the Kish and Scott method would make it possible to reselect the 9 units in $s_{g,1}$.

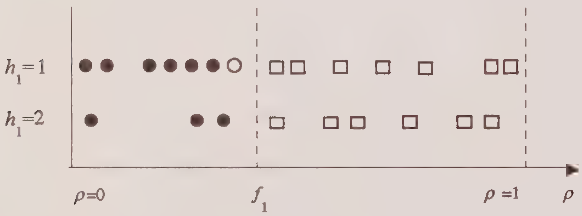


Figure 2. Coverage with method 2 (equidistant numbers).

We are in the same situation as in Figure (1), but this time the equidistant numbers ρ serve as the abscissa of the units. This equidistance is defined in each of the whole strata h_1 and the gaps we see in the sequence of numbers correspond to the units which are not in g. The first sample $s_{g,1}$ is composed of the units for which this number is less than the probability of inclusion f_1, regardless of the stratum. The second sample $s_{g,2}$ is composed of the 9 units with the smallest ρ and the coverage is 9, as with the Kish and Scott method (1971).

6. UPDATING BIRTHS AND DEATHS WITHIN STRATA

In this section and the following one, we consider the stratification (h) without reference to the period. The updating of births and deaths within strata is essentially a particular case of change of the strata of units. It is exactly as if the births entered the strata and the deaths left. We can therefore apply the previous methods. Let us take a look, in particular, at method 2.

In a stratum, the population $U_{h,t}$ of size $N_{h,t}$ varies with each updating carried out at time t. We will notate the births as $B_{h,t+1}$ and the deaths $D_{h,t+1}$ between t and as $t+1$, this yields $U_{h,t+1} = U_{h,t} + B_{h,t+1} - D_{h,t+1}$.

We consider the simple case where the probabilities of inclusion f_h remain uniform in $U_{h,t}$ and constant. The size $n_{h,t}$ of the sample $s_{h,t}$ is a rounded number to the integer of $N_{h,t} f_h$. The numbers $\rho_{i,t}$ change with each updating. Just before updating $s_{h,t}$, leading to $s_{h,t+1}$:

a) we make equidistant the numbers $\rho_{i,t-1}$ in $U_{h,t}$.

b) we attribute equidistant numbers to the units of $B_{h,t+1}$.

Let $\rho_{i,t}$ be the number so obtained. An initial solution would consist in selecting the $n_{h,t+1}$ units of $U_{h,t+1}$ with the smallest $\rho_{i,t}$. Note that these are no longer equidistant because we removed the deaths situated at random.

However, units with numbers close to f_h can leave the sample and then return on a future occasion. We remedy this by a rightward shift of the selection interval. Let ρ_{h,d_t} be the number of the beginning unit of the selection interval for $s_{h,t}$ and ρ_{h,e_t} that of the unit immediately following the end unit of this interval in $U_{h,t}$. In other words, the sample $s_{h,t}$ consists of the interval closed to the left and open to the right $[\rho_{h,d_t}, \rho_{h,e_t})$. Between t and $t+1$, the number of units of $U_{h,t+1}$ belonging to this interval becomes $m_{h,t+1}$. If $n_{h,t+1} \geq m_{h,t+1}$, the beginning of the interval for $s_{h,t+1}$ is fixed to the unit of number ρ_{h,d_t}, otherwise we shift the interval in such a way that its end is the unit of number ρ_{h,e_t}. We therefore have a slight involuntary rotation.

7. ROTATION BETWEEN TWO RESAMPLINGS

7.1 Rotation Without Updating of Births and Deaths

We can then stipulate a time of inclusion D_h whole and constant in the stratum. We have two variants, depending on whether we keep the same rounded number or vary it.

7.1.1 Fixed Rounded Number

We therefore have a size n_h strictly fixed during the rotation. We divide n_h into D_h whole numbers $n_{h,l}$, $(l = 1, \ldots, D_h)$ such that $|n_{h,l} - n_h/D_h| < 1$. Let q_h be the quotient and r_h the remainder of the division of $t - t_1$ by D_h and let $n_{h,0} = 0$. The sample at time t includes the units ranging from rank $1 + q_h n_h + \sum_{l=0}^{r_h} n_{h,l}$ to rank $(q_h + 1) n_h + \sum_{l=0}^{r_h} n_{h,l}$.

If $D_h = D$, we can stipulate in addition

$$\left|\sum_{h=1}^{H} n_{h,l} - \frac{n}{D}\right| < 1, l = 1, \ldots, D_h.$$

The variance of the rate of rotation is then practically nil.

However, the duration of inclusion is not controlled when $v_h < 1$: this yields $n_h = 0$ or $n_h = 1$. In the first case, there is no rotation, and in the second case, on the contrary, the time of exclusion can be considered too short. The following method makes it possible to obtain a rotation which corresponds to v_h.

7.1.2 Variable Rounded Number

The sample $s_{h,t}$ is defined based on the numbers rendered equidistant:

$$i \in s_{h,t} \Leftrightarrow \rho_{i,1} \in \left[f_h \frac{t - t_1}{D_h}, f_h \frac{t - t_1}{D_h} + f_h \right).$$

The sample size varies between $I(v_h)$ and $I(v_h) + 1$ in the stratum, and it is independent of the sizes in the other strata. This shows us what the result would be of the sample rotation advocated by Brewer *et al.* (1984) in the case of stratified sampling of fixed sized and uniform probability in each stratum.

7.2 Rotation With Updating of Births and Deaths

To simplify, we assume that each new survey wave is accompanied by the introduction of the births since the previous wave and a rotation. The method bifurcates into two procedures depending on whether or not we wish to respect exactly the durations of inclusion D_h between two resamplings.

7.2.1 Procedure A

The births are isolated in separate strata, and we wait for the resampling before subtracting the deaths. In this case each wave of births is dealt with exactly like an initial sampling after attributing the numbers ω_i. The sampling is carried out by stratifying with the same nomenclature (h), or with another more scattered or more confined. To simplify the notations, but without loss of generality, we assume that this is the same nomenclature. The index of stratification can then be written (b, h), where b crossed with h indicates the wave of births with a particular modality $b = 1$ corresponding to the units already existing during the first sampling or a previous resampling. This brings us back to the case of section 7.1 in each stratum (b, h) and the duration of inclusion is respected exactly.

The number of strata, and therefore of rounded numbers, is multiplied by the number of waves of births. The sample size can become fairly random with independent roundings (but less so than with Poisson sampling). It may therefore be worthwhile to link, at least partially, the rounded numbers. For example, we carry out a systematic rounding in the dimension h for each b or the reverse. We then keep these roundings and this is the 7.1.1 method which then applies rather than the 7.1.2 method.

7.2.2 Procedure B

In procedure B, we subtract the deaths at each survey wave. This is the type of updating presented in section 6. We would prefer a fixed duration of inclusion, but that is made difficult by the random number of deaths. At most, we can try to control a maximum duration of inclusion DM_h. We may also wish to prevent the units which have just left the sample from returning on a future occasion, which can occur if the rotation is slow. The idea is to get back to the algorithm described in section 6 by removing first of all from $s_{h,t}$ the units of which the previous duration of inclusion in $s_{h,t}$ attained DM_h. They are found the farthest to the left of the interval $[\rho_{h,d_t}, \rho_{h,e_t})$ and are mixed with the births too recent to have attained DM_h. However, these must still be removed in order for the distribution of the sample according to the generations to be correct. For that, it is sufficient to attribute to the births a fictitious previous duration of inclusion which falls between 1 and DM_h, just after defining the sample. For example, after defining $s_{h,t}$, we assign to each unit of $B_{h,t}$ belonging to the sample the same previous duration of inclusion in the sample as that of the unit of $U_{h,t-1}$ situated immediately to the left. Then let R_{h,d_t} be the highest rank among the ranks according to $\rho_{i,t}$ of the units of the interval associated with $s_{h,t}$ which have been included DM_h times in the sample; we discard the first units of $s_{h,t}$ up to and including rank R_{h,d_t}. Finally, this brings us back to the algorithm described in section 6 with, for ρ_{h,d_t}, the number of the unit of rank $R_{h,t} + 1, \rho_{h,e_t}$ remaining that of the unit which follows the unit of last rank in $s_{h,t}$.

8. RESAMPLING AFTER ROTATION

We now reselect the indices of strata h_1, h_2. We define the stratification $\mathbf{h}_1$ as a function of the procedure used for the updates of the births. With procedure A, we place the births in separate strata, this is the stratification defined by crossing the waves of births b with the nomenclature h_1. With procedure B, $\mathbf{h}_1$ is identical to h_1 However, we keep the notations of the independent quantities of b as f_{h_1}, D_{h_1}.

The selection of the new sample s_2, in a new stratification h_2 must be carried out at period $t = t_2$.

We begin by removing from the previous sample (at period $t = t_2 - 1$) the units which have attained the maximum authorized duration of inclusion. There remains a sample s'_1 of size $n'_{\mathbf{h}_1}$ of which we would like to conserve the maximum number of units in the resampling.

In the case without rotation examined in section 5, it was easy to define the resampling because the sample s_1 was composed of the units of lower rank according to ω_i in each stratum after a real number independent of the ω_i. In this instance, this number is 0. The resampling took place in the same manner by selecting the units of lower rank according to $\rho_{i,1}$, after this number, in the new strata.

After rotation this no longer works: there is no longer any real independent of the numbers such that the sample s'_1 is composed of units of lower rank after it. This is true even in the case where $f_{h_1} = f_1$. The problem is obviously aggravated with f_{h_1} varying by stratum. The idea which then comes to mind is to first carry out a transformation of the numbers in such a way that those from s'_1 find themselves at the beginning of [0, 1). This will then bring us back to the case without rotation. This is the same kind of idea which is presented by Hidiroglou, Choudhry and Lavallée (1991).

This transformation is fairly immediate in the particular case where the updates are done with procedure A and with the variable rounded number from section 7.1.2. Without

resampling, the selection interval at time t_2 would have been:

$$\rho_{i,1} \in \left[(t_2 - t_1) f_{h_1}/D_{h_1}, (t_2 - t_1) f_{h_1}/D_{h_1} + f_{h_1}\right).$$

The resampling results in new strata with probabilities f_{h_2}. These include the creations of units between the dates $t_2 - 1$ and t_2, to which we attribute equidistant numbers $\rho_{i,1}$, in each stratum h_2, independently of the survivors. They still contain units whose death has occurred since the previous sampling. It is possible to define a new sample s_2 in the same way as for Poisson sampling, by the interval, *i.e.*,

$$\rho_{i,1} \in \left[a_{h_1}, a_{h_1} + f_{h_2}\right),$$

where:

$$a_{h_1} = (t_2 - t_1) f_{h_1}/D_{h_1} + \max\left[0, f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2}\right].$$

Let us recall that we shift from the supplementary quantity

$$f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2}, \quad \text{if} \quad f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2} > 0,$$

to prevent the units which have just left the sample from returning too quickly.

As for Poisson sampling, the probability of a survivor being in the old and the new sample is then the maximum possible, namely:

$$\min\left(f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right), f_{h_2}\right).$$

However the size n'_{h_2} of this sample is random, whereas we want a sample of fixed size n_{h_2}. We obtain it by selecting, in each new stratum h_2, after having removed the deaths, the n_{h_2} units of lower rank according to $\eta_{i,1} = \rho_{i,1} - a_{h_1}$. This number therefore plays, for the resampling, the same role that ω_i played during the first sampling.

If, on the other hand, we chose procedure A with a fixed rounded number in the rotation or if we chose procedure B, we must begin again with the rank of the units of $\mathbf{h}_1$ during the last updating. This is the rank according to ω_i with procedure A or the rank according to $\rho_{t_2} - 1$ with procedure B. Let us assume that $N_{\mathbf{h}_1}$ is the size of the population at date $t_2 - 1$. Let $R_{\mathbf{h}_1, d}$ be the rank of the unit preceding the one of lower rank in s'_1 and $R_{\mathbf{h}_1}(i)$ the rank of unit i. The number used to classify the units in the new strata becomes:

$$\eta_{i,1} = \frac{R_{\mathbf{h}_1}(i) - 1 - a_{\mathbf{h}_1} + \delta_{\mathbf{h}_1}}{N_{\mathbf{h}_1}} \text{ modulo } 1,$$

where:

$$a_{\mathbf{h}_1} = R_{\mathbf{h}_1, d} + \max\left(0, n'_{\mathbf{h}_1}/N_{\mathbf{h}_1} - f_{h_2}\right).$$

With procedure A we can keep $\delta_{\mathbf{h}_1} = \theta_{\mathbf{h}_1}$ while we make a choice of $\delta_{\mathbf{h}_1}$ consistent with the last rounded number if procedure B is applied. However, because of the rotation, this choice has a minor impact on the coverage and it would be almost as well to select at random in [0, 1).

9. CONCLUSION

Algorithms based on equidistant numbers do not produce SRS. The first-order probabilities of inclusion are not exactly controlled and the second-order probabilities are unknown. During the changes of stratum, there remains a "trace" of the former strata in the new strata. The application of the SRS formulae to estimate the variance leads to biased results, generally in the direction of over-estimation. However, we think that the improvement in coverage during resamplings provided by the algorithms based on equidistant numbers outweighs the disadvantage of biased estimation of the variance and of the confidence intervals. According to section 5, the finer the stratification the greater this advantage. In particular, the use of equidistant numbers seems to be quite indicated with procedure A where the strata (b, h) are likely to be very small for the waves of births $(b > 1)$. The advantage of equidistant numbers is not as great with procedure B. However, making the numbers of births equidistant renders both the number of survivors reselected at each updating of the sample and the duration of inclusion less random.

However, let's take a quick look at what would change in the maintenance if we wanted to conserve SSRS. At each stage we must conserve the independent and uniform distribution of the ω_i. First of all, the phases of updating the births and of rotation between resamplings described in sections 6 and 7 apply while still conserving the same ω_i and the procedure is even simpler. The most delicate part is the resampling after the intermediate phase of rotation. The objective is to obtain not only a SSRS but also, if possible, the same coverage as for method 1 in section 5.

Let us assume that $\alpha_{\mathbf{h}_1}(j)$ is the number ω of the unit of rank j in a former stratum $\mathbf{h}_1$.

Let us assume first of all that, in a former stratum, all the units are such that $f_{h_2} \geq n'_{\mathbf{h}_1}/N_{\mathbf{h}_1}$. In particular, this occurs in all the strata for a sampling with a single rate in the sampled part, if we do not lower this rate. We then endeavour to find a transformation such that the numbers of the units of the sample are at the beginning of [0, 1). The simplest is the permutation:

$$\begin{cases} \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j + N_{\mathbf{h}_1} - R_{\mathbf{h}_1, d}), & j \leq R_{\mathbf{h}_1, d}, \\ \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j - R_{\mathbf{h}_1, d}), & j > R_{\mathbf{h}_1, d}. \end{cases}$$

However, a less costly transformation is:

$$\begin{cases}\beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j) + \alpha_{\mathbf{h}_1}(N_{\mathbf{h}_1}) - \alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d}), & j \leq R_{\mathbf{h}_1,d},\\ \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j) - \alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d}), & j > R_{\mathbf{h}_1,d}.\end{cases}$$

It is sufficient to find the result of $\alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d})$ and $\alpha_{\mathbf{h}_1}(N_{\mathbf{h}_1})$, after which a simple sequential calculation makes it possible to deduct β from α.

The Jacobian of the transformation is equal to 1 and consequently the numbers conserve their uniform distribution. Moreover, the joint distribution $p(s_1, s_2)$ is the same as if there had been no rotation. The demonstration is provided in Cotton and Hesse (1992, page 55). We therefore have the maximum coverage of SSRS.

If this yields units with $f_{h_2} < n'_{\mathbf{h}_1}/N_{\mathbf{h}_1}$ in the stratum and we apply the transformation, the units whose rank falls approximately between $N_{\mathbf{h}_1} f_{h_2}$ and $n'_{\mathbf{h}_1}$ are not reselected during the resampling but will be reintroduced during a future rotation. It is therefore preferable to use, for these units, a transformation which is situated just before f_{h_2} the new numbers. We must proceed by subsets according to the value of f_{h_2}. However, that tends to decrease the coverage.

ACKNOWLEDGEMENTS

The starting point of our work is an internal document from the Business Survey Methods Division of Statistics Canada: Hidiroglou, M.A., Srinath, K.P. (1990), Methods of integrated sampling for sub-annual business surveys.

We would like to thank a co-writer and an anonymous referee for their assistance in the drafting of this article.

Some of the methods proposed have been applied to the INSEE, but the opinions expressed are solely those of the authors.

APPENDIX
Probabilities of Inclusion in the Kish and Scott Method (1971)

Let us consider an example where the first-order probability of inclusion is not strictly controlled.

The population is divided into three parts A, B and C of equal size N. The first sampling is a SRS of $2a$ units in $A + B$ and a SRS of a units in C. During the second sampling, we wish to select a units in A and $2a$ units in $B + C$, while retaining the maximum number of units from the first sample and with uniform probability of inclusion a/N. The Kish and Scott method consists in adding or removing by SRS the appropriate number of units separately in A and in $B + C$. In A, the second marginal sampling is a SRS and the probability of inclusion is quite uniform. We will show that this is not the case in $B + C$. Let n_1 and n_2 be the sizes of the two successive samples in B. By symmetry, the probability of inclusion during the second sampling is uniform in B. It is equal to:

$$\begin{aligned}E(n_2)/N &= [E(n_1) + E(n_2 - n_1)]/N\\ &= a/N + E(n_2 - n_1)/N.\end{aligned}$$

If $n_1 = a, n_2 - n_1 = 0$; otherwise the expected value of $n_2 - n_1$ conditional on n_1 differs depending on the sign of $a - n_1$:

If $a - n_1 > 0, E[(n_2 - n_1)] \mid n_1] = (a - n_1)(N - n_1)/(2N - n_1 - a)$.

If $a - n_1 < 0, E[(n_2 - n_1) \mid n_1] = (a - n_1)n_1/(n_1 + a)$.

Note $p(n_1)$ the probability that the first sample will have the size n_1 in B. This yields:

$$E(n_2 - n_1) = \sum_{n_1} p(n_1) E[(n_2 - n_1) \mid n_1].$$

Since the sizes of A and B are equal, $p(n_1) = p(2a - n_1)$, therefore:

$$\begin{aligned}&E(n_2 - n_1)\\ &= \sum_{n_1<a} p(n_1)\{E[(n_2 - n_1) \mid n_1] + E[(n_2 - n_1) \mid (2a - n_1)]\}\\ &= \sum_{n_1<a} p(n_1)(a - n_1)[(N - n_1)/(2N - n_1 - a) - (2a - n_1)/(3a - n_1)]\\ &= (2a - N)\sum_{n_1<a} p(n_1)(a - n_1)^2/[(2N - n_1 - a)(3a - n_1)]\\ &= (2a - N)K, K > 0.\end{aligned}$$

Except in the case $2a - N = 0$, $E(n_2 - n_1)$ is not nil and $E(n_2)/N$ is different from a/N. The probability of inclusion is therefore not uniform in $B + C$.

REFERENCES

BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.

COTTON, F., and HESSE C. (1992). Tirages coordonnés d'échantillons. INSEE working paper E9206.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

HIDIROGLOU, M.A., CHOUDHRY, G.H., and LAVALLÉE, P. (1991). A sampling and estimation methodology for sub-annual business surveys. *Survey Methodology*, 17, 195-210.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 119-126
Statistics Canada

Empirical Bayes Estimation of Small Area Proportions Based on Ordinal Outcome Variables

PATRICK J. FARRELL[1]

ABSTRACT

Much research has been conducted into the modelling of ordinal responses. Some authors argue that, when the response variable is ordinal, inclusion of ordinality in the model to be estimated should improve model performance. Under the condition of ordinality, Campbell and Donner (1989) compared the asymptotic classification error rate of the multinominal logistic model to that of the ordinal logistic model of Anderson (1984). They showed that the ordinal logistic model had a lower expected asymptotic error rate than the multinominal logistic model. This paper also aims to compare the performance of ordinal and multinomial logistic models for ordinal responses. However, rather than focussing on classification efficiency, the assessment is made in the context of an application where the objective is to estimate small area proportions. More specifically, using multinominal and ordinal logistic models, the empirical Bayes approach proposed by Farrell, MacGibbon and Tomberlin (1997a) for estimating small area proportions based on binomial outcome data is extended to response variables consisting of more than two outcome categories. The properties of estimators based on these two models are compared via a simulation study in which the empirical Bayes methods proposed here are applied to data from the 1950 United States Census with the objective of predicting, for a small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level.

KEY WORDS: Bootstrap; Complex survey design; Logistic regression; Random effects models; Small area summary statistics; Taylor series.

1. INTRODUCTION

Much research has been conducted into the modelling of ordinal responses (see Albert and Chib 1993, Anderson 1984, Crouchley 1995, and McCullagh 1980). Some authors argue that, when the response variable is ordinal, inclusion of ordinality in the model to be estimated should improve model performance. Under the condition of ordinality, Campbell and Donner (1989) theoretically compared the asymptotic classification error rate of the multinomial logistic model to that of the ordinal logistic model of Anderson (1984), demonstrating that the ordinal model had a lower expected asymptotic error rate. However, in a subsequent simulation study, Campbell, Donner, and Webster (1991) illustrated that ordinal models classify less accurately than multinomial models under a variety of circumstances, and concluded that ordinal models confer no advantage when the main purpose of an analysis is classification.

This paper also aims to compare the performance of ordinal and multinomial logistic models for ordinal responses. However, rather than focussing on classification efficiency, the assessment is made in the context of an application where the objective is to estimate small area proportions.

The estimation of small area parameters is a finite population sampling problem which has received considerable attention. An excellent review of such research appears in Ghosh and Rao (1994). These authors demonstrate that as a compromise between synthetic and direct survey estimators, estimators based on empirical or hierarchical Bayes procedures are not subject to the large bias that is sometimes associated with a synthetic estimator (see Gonzales 1973), nor are they as variable as a direct survey estimator. A similar conclusion was drawn by Farrell, MacGibbon, and Tomberlin (1997a) in a study of the properties of an empirical Bayes estimator for small area proportions based on a binomial outcome variable.

Despite the numerous studies aimed at predicting small area proportions based on binomial response variables (see Dempster and Tomberlin 1980, MacGibbon and Tomberlin 1989, Farrell 1991, Farrell *et al.* 1997a, Malec, Sedransk, and Tompkins 1993, Stroud 1991, and Wong and Mason 1985), little attention has been given to estimating proportions based on response variables with more than two outcome categories. This paper extends the empirical Bayes approach of Farrell *et al.*, (1997a), to such response variables by basing the estimates on multinomial and ordinal logistic models. To compare the estimates of small area proportions based on an ordinal outcome variable using multinomial and ordinal models, the proposed empirical Bayes methods are applied to data from the 1950 United States Census in order to predict, for a given small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level.

For such an estimation problem, there are many issues which require attention. They include the selection of predictor variables for the model, model diagnostics, the sample design, and the properties of the estimators

[1] Patrick J. Farrell, Assistant Professor, Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia, B0P 1X0.

employed. For example, among the model diagnostics for the multinomial and ordinal models was an assessment of model fit which was based on residuals. For a description of this diagnostic and others, see Farrell (1991). The findings did not appear to indicate a lack of fit for either model. In this study, the focus is on investigating the properties of empirical Bayes estimators over repeated realizations of the sample design using a simulation. For many survey practitioners, such properties are of prime importance.

One concern associated with using an empirical Bayes estimation approach is that interval estimates do not attain the desired level of coverage, since the uncertainty that arises from having to estimate the parameters of the prior distribution is not accounted for. This study incorporates the suggestion of Laird and Louis (1987) to use bootstrap techniques for adjusting naive estimates of accuracy. Alternatively, Prasad and Rao (1990) have developed a procedure which attempts to account for the uncertainty not captured by the naive estimates. Although their approach was designed for three specific linear models containing random effects, Cressie (1992) has made certain conjectures as to when the procedure is appropriate. Of importance is the constraint that the outcome variable must follow a normal distribution.

The proposed empirical Bayes procedures based on multinomial and ordinal logistic models are presented in Section 2. The simulation study to compare multinomial and ordinal logistic models for ordinal responses is described in Section 3, while the conclusions and discussion are presented in Section 4.

2. ESTIMATION PROCEDURES

Consider a discrete small area characteristic of interest with M possible outcomes. The subscript m will reference these categories, where $m = 1, ..., M-1$ and $m^+ = 1, ..., M$. In addition, underlined lower case and capital letters will designate vectors, while bold capital letters will represent matrices.

The estimation procedures are illustrated under a two stage sample design, where individuals are sampled from selected local areas. Thus, local areas are the primary sampling units here. Let p_{im+} be the proportion of individuals in the i-th local area that belong to category m^+ of the response variable. Then

$$p_{im^+} = \sum_j y_{ijm^+}/N_i, \tag{2.1}$$

where y_{ijm+} is either zero or one, depending upon whether the j-th individual in local area i belongs to category m^+ of the characteristic of interest, and N_i is the population size of the i-th local area.

The approach employed by Farrell *et al.*, (1997a), to estimate small area proportions based on binomial outcome variables is extended here to allow for the estimation of p_{im+}. The procedure follows the explicitly model-based approach proposed by Dempster and Tomberlin (1980). Let π_{ijm+} represent the probability that the j-th individual within the i-th local area belongs to category m^+ of the response variable. Then, according to Royall (1970), p_{im+} in (2.1) is estimated by

$$\hat{p}_{im^+} = \left(\sum_{j \in S} y_{ijm^+} + \sum_{j \in S'} \hat{\pi}_{ijm^+} \right) \Big/ N_i, \tag{2.2}$$

where S is the set of n_i sampled individuals from local area i, and S' is the set of individuals in local area i not included in the sample. Values for the $\hat{\pi}_{ijm+}$ are required. To obtain these estimates, logistic regression models are used to describe the probabilities associated with individuals in the population.

Under a multinomial logistic model, the π_{ijm+} are described as follows:

$$\begin{gathered} \log(\pi_{ijm} / \pi_{ijM}) = \underline{X}_{ij}^T \underline{\beta}_m + \delta_{im}, \\ \underline{\delta}_i \sim \text{i.i.d. Normal } (\underline{0}, \boldsymbol{D}), \end{gathered} \tag{2.3}$$

where $\underline{\delta}_i^T = (\delta_{i1}, ..., \delta_{i(M-1)})$, $i = 1, ..., I$, and $\boldsymbol{D}$ is an unknown covariance matrix. In this model, $\underline{X}_{ij}$ is a vector of fixed effects predictor variables, the vector $\underline{\beta}_m$ contains the fixed effects parameters associated with the m-th category of the outcome variable of interest, and δ_{im} is a normally distributed random effect associated with the m-th category of the characteristic of interest in the i-th local area. The vector $\underline{X}_{ij}$ may include covariates at both the individual and aggregate levels. For sample designs of more than two stages, an analogous model would contain random effects for the sampling units at each stage, excluding the final one.

Note that the model in (2.3), unlike a similar model proposed by Malec *et al.*, (1993), does not contain interaction terms between the local area effects and the fixed effects predictor variables. However, terms to acknowledge such interaction could be included if they were deemed necessary.

To obtain Bayes estimates of the model parameters, values are assumed for the unknown parameters of the random effects distribution. Let $\underline{y}_{ij}^T = (y_{ij1}, ..., y_{ijM})$ be a vector for the ij-th sampled individual where the component associated with the category of the outcome variable to which the individual belongs has a value of one. The remaining entries are zero. If $\boldsymbol{Y}$ is a matrix with rows $\underline{y}_{ij}^T$, then the data are distributed as:

$$f(\boldsymbol{Y} \mid \underline{\beta}, \underline{\delta}_c) \, \alpha \prod_{ij} \pi_{ij1}^{y_{ij1}} \pi_{ij2}^{y_{ij2}} \ldots \pi_{ijM}^{y_{ijM}},$$

where $\underline{\beta}^T = (\underline{\beta}_1^T, ..., \underline{\beta}_{M-1}^T)$, and $\underline{\delta}_c^T = (\underline{\delta}_1^T, ..., \underline{\delta}_I^T)$. If a flat distribution is specified for the fixed effects, the distribution of the parameters is $f(\underline{\beta}, \underline{\delta}_c \mid \boldsymbol{D}_c) \propto \exp(-\frac{1}{2}\underline{\delta}_c^T \boldsymbol{D}_c \underline{\delta}_c)$, where $\boldsymbol{D}_c = \text{diag}(\boldsymbol{D}, \boldsymbol{D}, ..., \boldsymbol{D})$. The joint distribution of the data and the parameters is determined using $f(\boldsymbol{Y} \mid \underline{\beta}, \underline{\delta}_c)$ and $f(\underline{\beta}, \underline{\delta}_c \mid \boldsymbol{D}_c)$, and subsequently employed to obtain the posterior distribution of the parameters. Unfortunately, a

closed form for this posterior distribution cannot be derived due to the intractable integration required to obtain the marginal distribution of Y. A possible approach could be a stochastic integration method such as Gibbs sampling (see Zeger and Karim 1991). Ripley and Kirkland (1990) indicate that the drawbacks of such an approach include the intensive computations and questions about when the sampling process has achieved equilibrium. Since computing time is of particular concern for the simulation discussed in Section 3, this approach will not be pursued here. Alternatively, Breslow and Clayton (1993) state that there is still room for simple, approximate methods. Many authors have found that a multivariate normal approximation of the posterior works very well in practice (see Farrell *et al.* 1997a, Laird 1978, Tomberlin 1988, and Wong and Mason 1985). Breslow and Lin (1995) warn, however, that such an approach might yield inconsistent estimates for the fixed effects parameters. Thus, if $\hat{p}_{im+}$ is to be based on fixed effects estimates obtained in this manner, the same might apply to the consistency of $\hat{p}_{im+}$ as an estimator for p_{im+}.

Following Farrell *et al.* (1997a), the posterior distribution of the parameters is approximated as a multivariate normal distribution having its mean at the mode and covariance matrix equal to the inverse of the information matrix evaluated at the mode. The information matrix here is simply the second derivative of the posterior distribution taken with respect to $\underline{\beta}$ and $\underline{\delta}_c$. When values are specified for the unknown parameters of the random effects distribution, the resulting mode and covariance matrix constitute an initial set of estimates of the model parameters. Empirical Bayes estimates are then obtained by using the EM algorithm described by Dempster, Laird, and Rubin (1977) to determine estimates for the parameters of the random effects distribution. The algorithm converges quickly, taking only a few minutes in real time. For details on how the empirical Bayes estimates are obtained for a model based on a two stage sample design and a binomial response variable, see MacGibbon and Tomberlin (1989).

The empirical Bayes estimates of the model parameters are used in (2.2) to determine $\hat{p}_{im+}$. In developing an expression for the uncertainty of $\hat{p}_{im+}$, N_i is assumed to be known. Since the approach being used is model-based and predictive in nature, the uncertainty in $\hat{p}_{im+}$ arises solely from the $\sum \hat{\pi}_{ijm+}$ term; the $\sum y_{ijm+}$ term has zero variance. Thus, the mean square error of $\hat{p}_{im+}$ as a predictor for p_{im+} can be estimated as

$$\widehat{\mathrm{MSE}}(\hat{p}_{im+}) = \widehat{\mathrm{Var}}\left(\frac{\sum_{j \in S'} \hat{\pi}_{ijm+}}{N_i}\right) + \frac{\sum_{j \in S'} \hat{\pi}_{ijm+}(1 - \hat{\pi}_{ijm+})}{N_i^2}. \tag{2.4}$$

For sampled local areas, where n_i is greater than zero, the first term of (2.4) is of order $1/n_i$, while the second term is of order $1/N_i$. In this study, the approximation of the mean square error of $\hat{p}_{im+}$ is based on the first term only, which yields a useful approximation provided that N_i is large compared to n_i. For nonsampled local areas, the first term in (2.4) is of order 1; therefore it always dominates the second term.

To estimate the uncertainty of $\hat{p}_{im+}$, which is expressed as a non-linear function of the estimators of the fixed and random effects, the expression for $\hat{p}_{im+}$ is linearized by taking a first order multivariate Taylor series expansion about the realized values of the fixed and random effects. The variance of the resulting expression, call it $\widehat{\mathrm{Var}}(\hat{p}_{im+})$, is taken as an estimate of the uncertainty of $\hat{p}_{im+}$. Details of the Taylor series expansion are given in Farrell *et al.*, (1997a), for a binomial outcome variable.

When population micro-data for auxiliary variables are not available, $\hat{p}_{im+}$ in (2.2) cannot be determined. For non-linear models such as (2.3), prediction is not straightforward in this situation. However, an alternative estimator to $\hat{p}_{im+}$, say $\tilde{p}_{im+}$, which requires only local area summary statistics (a mean vector and finite population covariance matrix) for both continuous and categorical variables can be obtained by extending the approach proposed by Farrell, MacGibbon, and Tomberlin (1997b) for achieving this objective when estimating binomial small area parameters. The same Taylor series expansion that was used to estimate the accuracy of $\hat{p}_{im+}$ can be employed to obtain a measure of the uncertainty for $\tilde{p}_{im+}$, $\widehat{\mathrm{Var}}(\tilde{p}_{im+})$.

The approach described in this section can also be used to develop point and interval estimates for small area proportions based on $\hat{p}_{im+}$ and $\tilde{p}_{im+}$ when an ordinal model is used. In this study, a fixed and random effects model is proposed for the π_{ijm+} which is based on the ordinal model proposed by McCullagh (1980)

$$\log\left(\frac{\pi_{ij1} + \ldots + \pi_{ijm}}{\pi_{ij(m+1)} + \ldots + \pi_{ijm}}\right) = \beta_{0m} - \underline{X}_{ij}^T \underline{\beta} + \delta_{im}, \tag{2.5}$$

$$\underline{\delta}_i \sim \text{i.i.d. Normal } (\underline{0}, \boldsymbol{D}).$$

The vector $\underline{X}_{ij}$ contains the values of the fixed effects predictor variables for the ij-th individual, while $\underline{\beta}$ represents a vector of fixed effects parameters. Associated with the m-th category of the response variable is a constant term, β_{0m}. The random effects are again assumed to be normally distributed. Note that an important feature of the model in (2.5) is that the restriction $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ must hold in order for $\pi_{ij(m+1)} \geq 0$. A discussion concerning this constraint is given in Section 3.

The approach used to approximate the uncertainty in $\hat{p}_{im+}$ and $\tilde{p}_{im+}$ when π_{ijm+} is based on either (2.3) or (2.5) can be described as naive, since $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ and $\widehat{\mathrm{Var}}(\tilde{p}_{im+})$ do not account for the uncertainty which results from estimating the parameters of the random effects distribution. Thus, interval estimates for p_{im+} that are based on $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ and $\widehat{\mathrm{Var}}(\tilde{p}_{im+})$ are typically too short. Many approaches have been proposed for addressing this issue (see Carlin and Gelfand 1990, and Laird and Louis 1987). In this study, the Type III bootstrap proposed by Laird and Louis (1987) is used to adjust naively-estimated measures of uncertainty. The procedure is described in Farrell *et al.*,

(1997a), for a binomial outcome variable. It can be extended to (2.3) and (2.5), and is applicable regardless of whether estimation is based on $\hat{p}_{im+}$ or $\tilde{p}_{im+}$.

The procedure requires that a number of bootstrap samples, N_B, be generated from a given set of data. Suppose that small area estimation is to be based on $\hat{p}_{im+}$. For the b-th bootstrap sample, an estimate $\hat{p}_{bim+}$ for p_{im+} based on (2.3) or (2.5), along with a naive estimate of the variability of $\hat{p}_{bim+}$, $\widehat{\mathrm{Var}}(\hat{p}_{bim+})$ are obtained. The quantities $\hat{p}_{bim+}$ and $\widehat{\mathrm{Var}}(\hat{p}_{bim+})$ are determined for each of N_B bootstrap samples, and used to calculate a bootstrap-adjusted estimate of the variability associated with $\hat{p}_{im+}$:

$$\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im^+}) = \frac{\sum_b \widehat{\mathrm{Var}}(\hat{p}_{bim^+})}{N_B} + \frac{\sum_b (\hat{p}_{bim^+} - \hat{p}^{\{B\}}_{im^+})^2}{N_B - 1},$$

$$\text{where} \quad \hat{p}^{\{B\}}_{im^+} = \frac{\sum_b \hat{p}_{bim^+}}{N_B}.$$

Note that even though individuals are not selected by simple random sampling without replacement in this study, survey weights have not been attached to the records. However, in practice, the weights attached to a record will vary due to features of the survey design, such as differential nonresponse and clustering. In this study, the models account for the effects of these features. Further research is necessary to determine what impact the incorporation of survey weights into the models would have on the bootstrapping procedure.

3. A DATA EXAMPLE

A comparison of the estimates for small area proportions based on multinomial and ordinal logistic models was carried out using a simulation study where the response variable was ordinal. The data set is based on a 1% sample of the 1950 United States Census (United States Bureau of the Census 1984). Data based on the 1950 Census is used since it constitutes a public use microdata sample, and none of the more recent census data is available in this form. Thus, the results below for the multinomial and ordinal models are obtained by using predictor variable data for each individual within a local area. For a discussion of the difficulties encountered in obtaining microdata, see Bethlehem, Keller, and Pannekoek (1990).

The application considered is the estimation of the proportion of individuals in a given local area associated with each of the three categories of an ordinal outcome variable representing total personal income, where a local area is typically specified to be a state. This variable encompasses all sources of income, including wages and salaries, business income, and net income from other sources. An individual is regarded as having a low (less than \$2,500), medium (\$2,500 to under \$10,000) or high (\$10,000 and over) level of total personal income during 1949. Thus, $m = 1$ for low income (Category 1), $m = 2$ for medium income (Category 2), and $m = 3$ for high income (Category 3). The multinomial and ordinal models were each used to obtain point and interval estimates for 42 local areas. Twenty of these areas were sampled, the others were not. Note that individuals with no income were included in Category 1. An alternative approach would have been a two stage model; a first stage logistic model for the probability of non-zero income, and a second stage multinomial or ordinal model for income category conditional on non-zero income.

In practice, historical data are often available for survey planning purposes. For example, variable selection for purposes of model predictions could be based on previous census data. To emulate this situation, a random sample of size 2,000 was selected from the 1% sample. Variables for model prediction were determined by applying a stepwise logistic regression procedure. The variables selected were age, gender, and race. With regards to race, individuals were categorized as white, negro, or other.

Thus, the multinomial and ordinal models used in this study included four individual level predictor variables for age, gender, and race (two indicator variables were required to code the various races). However, they also contained four local area variables representing average age, the proportion of males, the proportion of whites, and the proportion of negroes. Regardless of which model is considered, these local area variables are necessary since, when they are excluded, a relationship is noted between the expected value of $\hat{p}_{im+}$ and its bias, where as the expected value increases, the bias increases from large negative to large positive values. The inclusion of domain level covariates removes this correlation. Therefore, since local area variables are also included in the models, the multinomial model contains eighteen fixed effects parameters (two for each of the individual level and local area predictor variables, and two constant terms) and forty random effects (two for each of the twenty sampled local areas), while the ordinal model contains ten fixed effects parameters (one for each of the individual level and local area predictor variables, and two constant terms) and forty random effects (two for each of the twenty sampled local areas). For a detailed study comparing logistic regression models for estimating small area proportions with and without domain level covariates which uses binomial outcome data, see Farrell *et al.*, (1997a).

The data for estimating the proportions of individuals in each local area belonging to the various income level categories were obtained from the 1% sample using a self-weighting two stage sample design. In the first stage, 20 out of 42 local areas were selected, without replacement, using probabilities proportional to size (PPS). More specifically, the approach used to select these local areas was randomized systematic selection of primary sampling units with PPS (see Kish 1965, p. 230). Then, at the second stage, 50 individuals were randomly selected from each

chosen local area. A total of 500 samples were drawn using this two stage design; however, resampling was not performed at the local area selection stage. Thus, the same 20 local areas were sampled in each of the 500 replicates. For these 20 sampled local areas, the average local area proportions for Categories 1, 2, and 3 of income level are 0.7142, 0.2260, and 0.0598.

Note that for the ordinal model, the constraint $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ must hold in order for $\pi_{ij2} \geq 0$. A check of this constraint for each of the 500 samples using the estimates for the constant terms and the random effects indicated that it held at all times. In fact, it was discovered that in each of the 500 samples taken, the difference in the estimates for the constant terms was always positive, at least two orders of magnitude larger than the majority of the absolute differences of the random effects estimates, and always one order of magnitude bigger. Thus, the constant terms in the model dominate over the random effects.

To compare the properties of estimators for small area proportions over repeated realizations of the sample design, for each of the 500 samples selected the quantities $\hat{p}_{im+}$, $\widehat{\mathrm{Var}}(\hat{p}_{im+})$, and $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ associated with each income level category were obtained for each local area, sampled or not, using both the multinomial and ordinal models. For each model, the estimates for $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ and $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ were used to construct naive and bootstrap-adjusted empirical Bayes symmetric 95% confidence intervals, respectively. Estimates for $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ were obtained by using the bootstrap procedure to generate 100 bootstrap samples from each of the 500 simulation samples.

Note that for the ordinal model, the constraint $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ must also hold in the bootstrap procedure for random effects generated from an estimated distribution; otherwise negative estimates for some of the probabilities π_{ijm+} will result when creating bootstrap samples. Over the course of the simulation for the application considered here, no negative probabilities were encountered when bootstrapping. One approach for assessing the likelihood of negative probabilities during the bootstrap procedure is to consider the ratio of the difference $\hat{\beta}_{02} - \hat{\beta}_{01}$ to the estimated prior standard deviation of the difference $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. This ratio was determined for each sampled local area in each of the 500 simulation samples taken. The average of this entire set of ratios was 6.8, and none were found to be less than 5.8. Thus, the difference $\hat{\beta}_{02} - \hat{\beta}_{01}$ was determined to always be at least 5.8 times the estimated standard deviation of the difference $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. Based on the empirical rule, a rule of thumb would be to conclude that when the ratio described above is at least three, it is highly unlikely that negative probabilities will arise when bootstrapping.

Table 1 presents average summary statistics over the 500 simulation samples obtained for the multinomial and ordinal models across all sampled local areas for each of three income level categories. A study of the stability of these statistics was conducted by investigating how they changed as additional samples were taken. Only slight changes were observed once 150 samples had been reached. Table 1 includes the summary statistics obtained for the first 200 samples in brackets for comparative purposes.

For each income category, two summary statistics shown in Table 1 were evaluated to compare the design bias of $\hat{p}_{im+}$ for the multinomial and ordinal models; the average bias of $\hat{p}_{im+}$, and the average absolute bias of $\hat{p}_{im+}$. The average bias is simply the mean over all sampled local areas of the differences obtained when the actual proportion, p_{im+}, for the i-th local area is subtracted from the average point estimate for the area over the 500 simulation samples. The average absolute bias is defined similarly, except that the absolute value of each difference is used. Generally speaking, the results obtained for these two summary statistics were slightly better for the ordinal model, regardless of the income category considered. However, the multinomial model did result in a somewhat smaller average bias for $\hat{p}_{im+}$ for the low income category.

For each sampled local area, empirical root mean square errors (RMSE's) were computed over the 500 simulation samples under each model for the three income categories. For each model and income level combination, the appropriate empirical RMSE's were averaged over all sampled local areas, resulting in the average empirical RMSE's presented in Table 1. Once again, the performance of the ordinal model is slightly better for all three income level categories.

To study the reduction in empirical RMSE when a model-based approach to estimation is used instead of a classical design unbiased method, average empirical RMSE's analogous to those in Table 1 based on the 500 samples were computed using the observed local area sample proportions in place of $\hat{p}_{im+}$. The average empirical RMSE's obtained were substantially larger (0.0617, 0.0564, and 0.0311 for the low, medium, and high income level categories) than those based on $\hat{p}_{im+}$ under either model.

Table 1 also includes summary statistics over all sampled local areas which relate naive and bootstrap measures of variability in $\hat{p}_{im+}$ to average empirical RMSE. For each income level category, the average relative bias and the average absolute relative bias of the square root of $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ as an estimate of empirical RMSE are shown in Table 1 for the multinomial and ordinal models. The average relative bias is simply the mean over all sampled local areas of the values obtained when the difference resulting from the subtraction of the empirical RMSE for the i-th local area from the average of the square root of $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ for the area over the 500 simulation samples is divided by the empirical RMSE. The average absolute bias is defined similarly, expect that the absolute value of each difference is used. The table also presents similar averages for the bootstrap-adjusted measures of variability, $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$. For both the multinomial and ordinal logistic models, the average relative bias and average absolute relative bias of the bootstrap-adjusted estimates of variability are substantially smaller in magnitude than their naive counterparts for all three income level categories. In

Table 1
Average Summary Statistics based on 500 Simulation Samples for the Multinomial and Ordinal Logistic Models across all Sampled Local Areas for each Income Level Category.
The average summary statistics obtained over the first 200 simulation samples are included in brackets for comparative purposes

Average	Low Income Level		Medium Income Level		High Income Level	
	Multinomial	Ordinal	Multinomial	Ordinal	Multinomial	Ordinal
Bias of $\hat{p}_{im+}$	-0.0004 (-0.0004)	-0.0005 (-0.0006)	-0.0007 (-0.0006)	-0.0004 (-0.0003)	0.0011 (0.0010)	0.0009 (0.0009)
Absolute Bias of $\hat{p}_{im+}$	0.0076 (0.0078)	0.0051 (0.0055)	0.0089 (0.0085)	0.0048 (0.0046)	0.0108 (0.0106)	0.0074 (0.0073)
Empirical RMSE	0.0479 (0.0483)	0.0467 (0.0469)	0.0417 (0.0414)	0.0401 (0.0402)	0.0236 (0.0233)	0.0231 (0.0229)
Relative Bias of $\sqrt{\widehat{\mathrm{Var}}(\hat{p}_{im+})}$	-0.1192 (-0.1197)	-0.1125 (-0.1128)	-0.1273 (-0.1276)	-0.1180 (-0.1186)	-0.1524 (-0.1521)	-0.1376 (-0.1372)
Absolute Relative Bias of $\sqrt{\widehat{\mathrm{Var}}(\hat{p}_{im+})}$	0.1192 (0.1197)	0.1125 (0.1128)	0.1273 (0.1276)	0.1180 (0.1186)	0.1524 (0.1521)	0.1376 (0.1372)
Relative Bias of $\sqrt{\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})}$	-0.0275 (-0.0272)	-0.0173 (-0.0175)	-0.0309 (-0.0314)	-0.0204 (-0.0207)	-0.0391 (-0.0393)	-0.0273 (-0.0269)
Absolute Relative Bias of $\sqrt{\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})}$	0.0294 (0.0290)	0.0227 (0.0228)	0.0349 (0.0343)	0.0263 (0.0265)	0.0450 (0.0446)	0.0353 (0.0347)
Naive Coverage Rate	91.35 (91.325)	91.91 (91.875)	91.19 (91.225)	91.78 (91.750)	90.67 (90.650)	91.26 (91.300)
Absolute Deviation of Naive Coverage from the 95% Nominal Rate	3.65 (3.675)	3.09 (3.125)	3.81 (3.775)	3.22 (3.250)	4.33 (4.350)	3.74 (3.700)
Adjusted Coverage Rate	94.44 (94.400)	94.75 (94.775)	94.37 (94.350)	94.68 (94.650)	93.91 (93.925)	94.40 (94.375)
Absolute Deviation of Adjusted Coverage from the 95% Nominal Rate	1.58 (1.600)	1.43 (1.425)	1.71 (1.725)	1.50 (1.525)	1.91 (1.900)	1.62 (1.650)

addition, these bootstrap-adjusted average summary statistics are all very small, which indicates that the bootstrap-adjusted estimates of variability are capable of incorporating most of the uncertainty that arises from having to estimate the distribution of the random effects.

For each sampled local area, naive and bootstrap-adjusted coverage rates based on 95% interval estimates were computed over the 500 samples under each model for the three income level categories. Over all income level and model combinations, the bootstrap-adjusted coverage rates for individual local areas ranged from 92.2% to 97.6%. Since an approximate bound for the Monte Carlo error is 3 $\sqrt{(0.95)(0.05)/500}$, or 0.029, all bootstrap-adjusted coverage rates are within 3 standard errors of 95%.

For each model and income level combination, the appropriate coverage rates were averaged over all sampled local areas, resulting in the average naive and bootstrap-adjusted coverage rates in Table 1. A number of observations can be made which hold for each income level category. For both multinomial and ordinal models, the average coverage rates for the bootstrap-adjusted intervals are much closer to the 95% nominal rate than those associated with the naive intervals. However, both the average naive and bootstrap-adjusted coverage rates for the ordinal model are slightly better than counterparts for the multinomial model. This is also the case for the average absolute deviation of both the naive and bootstrap-adjusted coverage rates from the 95% nominal rate. The average absolute deviation of the naive coverage rates from the 95% nominal rate is simply the mean over all sampled local areas of the absolute values of the differences obtained when the 95% nominal rate is subtracted from the naive coverage rates for the sampled local areas over the 500 simulation samples. The average absolute deviation of the bootstrap-adjusted coverage rates from the 95% nominal rate is defined analogously.

Twenty-two local areas were not sampled. Estimates for the proportion of individuals associated with each income level category were also obtained for these areas using the multinomial and ordinal models. The findings were similar to those for sampled local areas. However, the performance of the models deteriorated somewhat, since nonsampled local areas constitute a holdout sample. For a detailed evaluation of results associated with nonsampled local areas, see Farrell *et al.* (1997a).

A comparison of the estimates for the three income level categories based on micro-data, $\hat{p}_{im+}$, with those based on local area summary statistics, $\tilde{p}_{im+}$, was also made for each

model. For both models, the results obtained for $\tilde{p}_{im+}$ were gratifyingly close to those obtained using $\hat{p}_{im+}$, although those obtained for $\hat{p}_{im+}$ were slightly better. Similar findings were obtained by Farrell *et al.*, (1997b) in a detailed comparison of $\hat{p}_{im+}$ and $\tilde{p}_{im+}$ for a binomial outcome variable.

4. CONCLUSION

Using multinomial and ordinal logistic models, the empirical Bayes approach proposed by Farrell *et al.*, (1997a), for estimating small area proportions based on binomial outcome data has been extended to accommodate outcome variables with more than two categories. It was found that the performance of the approach is preserved for multicategorical outcome data.

To compare the estimates of small area proportions based on an ordinal outcome variable using multinomial and ordinal logistic models, the proposed empirical Bayes methods based on these two models were applied to data from the 1950 United States Census with the objective of predicting, for a small area, the proportion of individuals who belong to the various categories of an ordinal response variable representing income level. The estimates based on the ordinal model were only slightly better in terms of design bias, empirical RMSE, and coverage rates. In addition, an important feature of the ordinal logistic model is that the constraint $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ must hold in order for $\pi_{ij(m+1)} \geq 0$. Since the results for the multinomial and ordinal models in the simulation were very similar, a multinomial model could be used for estimating small area proportions based on ordinal outcome variables when there is concern that fitting an ordinal model may result in negative estimates for some of these probabilities.

ACKNOWLEDGEMENTS

This research was supported by NSERC of Canada. The author is grateful to the associate editor and the referees for their valuable comments and suggestions.

REFERENCES

ALBERT, J.H., and CHIB, S. (1993). Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association*, 88, 669-679.

ANDERSON, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, 46, 1-30.

BETHLEHEM, J.G., KELLER, W.J., and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

BRESLOW, N.E., and CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

BRESLOW, N.E., and LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82, 81-91.

CAMPBELL, M.K., and DONNER, A. (1989). Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association*, 84, 587-591.

CAMPBELL, M.K., DONNER, A., and WEBSTER, K.M. (1991). Are ordinal models useful for classification? *Statistics in Medicine*, 10, 383-394.

CARLIN, B.P., and GELFAND, A.E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85, 105-114.

CRESSIE, N. (1992). REML Estimation in empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

CROUCHLEY, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, 90, 489-498.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

DEMPSTER, A.P., and TOMBERLIN, T.J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.

FARRELL, P.J. (1991). Empirical Bayes Estimation of Small Area Proportions. PhD. dissertation, Department of Management Science, McGill University, Montreal, Quebec, Canada.

FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997a). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997b). Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

GONZALES, M.E. (1973). Use and evaluation of synthetic estimation. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons Inc.

LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.

LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.

MacGIBBON, B., and TOMBERLIN, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237-252.

MALEC, D., SEDRANSK, J., and TOMPKINS, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. In *Case Studies in Bayesian Statistics*, (Eds. C. Gatsonis, J.S. Hodges, R. Kasf, and N.D. Singpurwalla). New York: Springer Verlag.

McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109-142.

PRASAD, N.G.N., and RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

RIPLEY, B.D., and KIRKLAND, M.D. (1990). Iterative simulation methods. *Journal of Computational and Applied Mathematics*, 31, 165-172.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 74, 1-12.

STROUD, T.W.F. (1991). Hierarchical Bayes predictive means and variances with application to sample survey inference. *Communications in Statistics, Theory and Methods*, 20, 13-36.

TOMBERLIN, T.J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.

UNITED STATES BUREAU OF THE CENSUS (1984). Census of the Population, 1950: Public Use Microdata Sample Technical Documentation, edited by J.G. Keane, Washington, D.C.

WONG, G.Y., and MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

ZEGER, S.L., and KARIM, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79-86.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 127-135
Statistics Canada

Poststratification Into Many Categories Using Hierarchical Logistic Regression

ANDREW GELMAN and THOMAS C. LITTLE[1]

ABSTRACT

A standard method for correcting for unequal sampling probabilities and nonresponse in sample surveys is poststratification: that is, dividing the population into several categories, estimating the distribution of responses in each category, and then counting each category in proportion to its size in the population. We consider poststratification as a general framework that includes many weighting schemes used in survey analysis (see Little 1993). We construct a hierarchical logistic regression model for the mean of a binary response variable conditional on poststratification cells. The hierarchical model allows us to fit many more cells than is possible using classical methods, and thus to include much more population-level information, while at the same time including all the information used in standard survey sampling inferences. We are thus combining the modeling approach often used in small-area estimation with the population information used in poststratification. We apply the method to a set of U.S. pre-election polls, poststratified by state as well as the usual demographic variables. We evaluate the models graphically by comparing to state-level election outcomes.

KEY WORDS: Bayesian inference; Election forecasting; Nonresponse; Opinion polls; Sample surveys.

1. INTRODUCTION

It is standard practice for weighting in opinion polls to be based entirely or primarily on poststratification, which we use generally to refer to any estimation scheme that adjusts to population totals. The basic approach is to divide the population into a number of categories, within each of which the survey is analyzed as simple random sampling. The poststratification step is to estimate population quantities by averaging estimates in the categories, counting each category in proportion to its size in the population. Poststratification categories are typically based on demographic characteristics (sex, age, *etc.*) as well as any variables used in stratification. Another level of complication, which we do not address here, would occur under cluster sampling.

There is a fundamental difficulty in setting up poststratification categories. It is desirable to divide the population into many small categories in order for the assumption of simple random sampling within categories to be reasonable. But if the number of respondents per category is small, it is difficult to accurately estimate the average response within each category. For example, if we poststratify by sex, ethnicity, age, education, and region of the U.S., some cells may be empty in the sample, whereas others may have only one or two respondents.

A general solution to this problem is to model the responses conditional on the poststratification variables (see Little 1993). For example, the standard approach to adjusting for several demographic variables is to rake across one-way or two-way margins (*i.e.*, iterative proportional fitting, Deming and Stephan 1940), which essentially corresponds to poststratification on the complete multi-way table, but with a model of the responses, conditional on the demographic variables, that sets higher-level interactions to zero. Methods based on smoothing weights can also be viewed as poststratification, with corresponding models on the responses (see Little 1991). When the poststratification categories follow a hierarchical structure (for example, persons within states in the U.S.), one can improve efficiency of estimation by fitting a hierarchical model (*e.g.*, Lazzeroni and Little 1997). In the related context of regression estimation, Longford (1996) demonstrates the potential for hierarchical linear models to improve the precision of small area estimates based on sample survey data.

In this paper, we set up a hierarchical logistic regression model to be used for poststratification estimates for a binary variable. The advantage of the model, compared to standard poststratification, is that it allows for the use of many more categories, and thus much more detailed population information. The practical gains from this method are greatest for small subgroups of the population. We apply the method to the state-level results of a set of U.S. pre-election polls. This example has the nice feature that we can check our inferences externally by comparing to state-level election outcomes. Details appear in an appendix for computing the hierarchical model using an approximate EM algorithm.

2. MODEL

2.1 Sampling and Poststratification Information

Consider a partition of the population into R categorical variables, where the r-th variable has J_r levels, for a total of $J = \Pi_{r=1}^{R} J_r$ categories (cells), which we label $j = 1, ..., J$.

[1] Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027 and Thomas C. Little, Morgan Stanley Dean Witter, New York, NY.

Assume that N_j, the number of units in the population in category j, is known for all j. Let y be a binary response of interest; label the population mean response in each category j as π_j. Then the overall population mean is $\bar{Y} = \sum_j N_j \pi_j / \sum_j N_j$. Assume that the population is large enough that we can ignore all finite-population corrections.

A sample survey is now conducted in order to estimate $\bar{Y}$ (and perhaps some other combinations of the π_j's). For each j, let n_j be the number of units in category j in the sample. Conditional on the R explanatory variables, assume that nonresponse is ignorable (Rubin 1976). Thus, the R variables should include all information used to construct survey weights, as well as any other variables that might be informative about y.

For the example we shall consider in Section 3, we categorize the population of adults in the 48 contiguous U.S. states by $R = 5$ variables: state of residence, sex, ethnicity, age, and education, with $(J_1, ..., J_5) = (48, 2, 2, 4, 4)$. (Ethnicity, age, and education are discretized into 4 categories each, as described in Section 3.1.) The $J = 3{,}072$ categories range from "Alabama, male, black, 18-29, not high school graduate" to "Wyoming, female, nonblack, 65+, college graduate," and, from the U.S. Census, we have good estimates of N_j in each of these categories. We shall consider population estimates (summing over all 3,072 categories) and also estimates within individual states (separately summing over the 64 categories for each state). It is impossible for a reasonably-sized sample survey to allow independent estimates of the mean responses π_j for each category j (in fact, the vast majority of categories will be empty or contain just one respondent), and so it is necessary to model the π_j's in order to poststratify and thus make use of the known category sizes N_j. The (potential) advantage of poststratification is to correct for differential nonresponse rates among the categories.

2.2 Regression Modelling in the Context of Poststratification

One can set up a logistic regression model for the probability π_j of a "yes" for respondents in category j:

$$\text{logit}(\pi_j) = X_j \beta, \tag{1}$$

where X is a matrix of indicator variables, and X_j is the j-th row of X. If we were to assume a uniform prior distribution on β, then Bayesian inference, for different choices of X, under this model corresponds closely to various classical weighting schemes. These correspondences, which we present below, are general and rely on the linearity of the assumed model (that is, $X_j\beta$ in (1)). (In the case of binary data, which we are considering in this paper, the classical and uniform-prior-Bayesian estimates are not identical, because of the nonlinear logistic transformation in (1), but for large samples the differences are minor.)

The following models correspond to the most commonly used classical poststratification estimates.

- Setting X to the $J \times J$ identity matrix corresponds to weighting each unit in cell j by N_j/n_j; that is, simple poststratification. This method is well known to work well only if the n_j's are reasonably large (and it will not work at all if $n_j = 0$ for any j).
- If we set X to the $J \times (\sum_{r=1}^{R} J_r)$ matrix of indicators for each individual variable, then the estimate of $\bar{Y}$ corresponds approximately to that obtained by raking across all R one-way margins.
- Including various interactions in X corresponds to including these same interactions in the raking. To put it most generally, assuming "structure" of any kind in X corresponds to pooling the poststratification across cells in some way.
- Including no explanatory variables in the model (that is letting X be simply a vector of 1's) leads to the sample mean estimate $\bar{y}$.

See Holt and Smith (1979) and Little (1993) for more discussion of the relation between weighting estimates and poststratification.

2.3 Hierarchical Regression Modelling for Partial Pooling

When the number of cells is large, none of the above options makes efficient use of the information provided by the categories (for example, simple poststratification gives estimates that are too variable, but if we exclude explanatory variables with many categories, we are discarding important information). Instead, we allow partial pooling across cells by setting up a mixed-effects model (see, *e.g.*, Clayton 1996). We write the vector β as $(\alpha, \gamma_1, ..., \gamma_L)$, where α is a subvector of unpooled coefficients and each γ_l, for $l = 1, ..., L$, is a subvector of coefficients (γ_{kl}) to which we fit a hierarchical model:

$$\gamma_{kl} \overset{\text{ind}}{\sim} N(0, \tau_l^2), k = 1, ..., K_l.$$

Setting τ_l to zero corresponds to excluding a set of variables; setting τ_l to ∞ corresponds to a noninformative prior distribution on the γ_{kl} parameters.

Given the responses y_i in categories j, we construct an $n \times J$ categorization matrix C, for which $C_{ij} = 1$ if respondent i is in cell j. Let $Z = CX$. The model (1) then can be written in the standard form of a hierarchical logistic regression model as

$$y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = Z\beta$$
$$\beta \sim N(0, \textstyle\sum_\beta),$$

where $\sum_\beta^{-1}$ is a diagonal matrix with 0 for each element of α, followed by τ_l^{-2} for each element of γ_l, for each l. We use the notation p_i, for the probability corresponding to the unit i, as distinguished from π_j, the aggregate probability corresponding to the category j. See Nordberg (1989) and

Belin, Diffendal, Mack, Rubin, Schafer and Zaslavsky (1993) for general discussions of hierarchical logistic regression models for survey data.

2.4 Inference Under the Model

To perform inferences about population quantities, we use the following empirical Bayes strategy: first, estimate the hyperparameters τ_l, given the data y; second, perform Bayesian inference for the regression coefficients β, given y and the estimated τ_l's; third, compute inferences for the vector of cell means $\pi = \text{logit}^{-1}(X\beta)$; fourth, compute inferences for population quantities by summing $N_j\pi_j$'s. We view this approach as an approximation to the full Bayesian analysis, which averages over the parameters τ_l. The two approaches will differ the most when components τ_l are imprecisely estimated or are indistinguishable from 0 (see for example, Gelman, Carlin, Stern and Rubin (1995), Section 5.5). In the example we consider here, this is not a problem because the various components are clearly estimated to be different from 0. If this were not the case, it would probably be worth putting in the additional programming effort for a full Bayes analysis. The focus of this paper, however, is on the effectiveness of combining hierarchical modeling with poststratification, not on the relatively minor technical differences between Bayes and empirical Bayes analyses.

The shrinkage of the cell estimates comes in the second step, and the amount of shrinkage depends both on the sample sizes n_j and the data $\bar{y}_{.j}$. More shrinkage occurs for smaller values of n_j and for values of $\bar{y}_{.j}$ far from the predictions based on the logistic regression model. In addition, more shrinkage occurs if the parameters τ_l are small. A batch of coefficients γ_l with little predictive power will be shrunk toward zero in the estimation, because τ_l will be estimated to have a small value. This is how we can include a large number of coefficients in the hierarchical model without the estimates of population quantities becoming too variable.

3. APPLICATION: BREAKING DOWN NATIONAL SURVEYS BY STATE

3.1 Survey Data

We apply the above methodology to state-by-state results from seven national opinion polls of registered voters conducted by the CBS television network during the two weeks immediately preceding the 1988 U.S. Presidential election. To follow our general notation, we assign $y_i = 1$ to supporters of Bush and $y_i = 0$ to supporters of Dukakis; we discard the respondents who expressed no opinion (about 15% of the total; we follow standard practice and count respondents who "lean" toward one of the candidates as full supporters). Since no data were collected from Hawaii and Alaska, only the 48 contiguous states are included in the model. Washington, D.C., although included in the surveys, was excluded from this analysis because its voting preferences are so different from the other states that a generalized linear model that fit the 48 states would not fit D.C. well, and as a result, the data from D.C. would unduly influence the results for the states. Since there are few observations for the smaller states and the between-poll variation in the estimated support for Bush is within binomial sampling variability (as measured by a χ^2 test of equality of the proportions of support for Bush in the seven polls), we combine the data from all the polls.

CBS creates survey weights by raking on the following variables, with default classifications for item nonresponse shown in brackets:

Census region:	Northeast, South, North Central, West
sex:	male, female
ethnicity:	black, [white/other]
age:	18-29, 30-44, [45-64], 65+
education:	not high school grad, [high school grad], some college, college grad.

The raking includes all main effects plus the interactions of sex × ethnicity and age × education. We include all these variables as fixed effects in our logistic regression model, excluding from our analysis the relatively few respondents with nonresponse in any of the demographic variables. The CBS weights also correct for number of telephone lines and number of adults in household, which affect sampling probabilities; these have minor effects on estimates for Presidential preference (see Little 1996, chapter 3), and we do not include them in our model. Further details of the CBS survey methodology and adjustment appear in Voss, Gelman, and King (1995).

Our model goes beyond the CBS analysis by including indicators for the 48 states as random effects, clustered into four batches corresponding to the four census regions. We check the performance of the model by comparing estimates for each state to the observed Presidential election. (Opinion polls just before the election are reliable indicators of the actual election outcome; see, *e.g.*, Gelman and King 1993.) We also compare the stability of estimates based on different polls over a short period of time.

3.2 Population Data for Poststratification

In order to poststratify on all the variables listed above, along with state, we need the joint population distribution of the demographic variables within each state: that is, population totals N_j for each of the $2 \times 2 \times 4 \times 48$ cells of sex × ethnicity × age × state. Since the target population is registered voters, we should use the population distribution of registered voters. As an approximation to that distribution we use the crosstabulations available in the Public Use Micro Survey (PUMS) data for all citizens of age 18 and over. The PUMS data contain records for 5% of the housing units in the U.S. and the persons in them, including over 12 million persons and over 5 million housing units. These data are a stratified sample of the approximately 15.9% of housing units that received long-form questionnaires in the 1990 Census. Persons in

institutions and other group quarters are also included in the sample. Weights are given for both the housing unit and persons within the unit based on sampling probabilities and adjustment to Census totals for variables included in the short-form questionnaire. We use the weighted PUMS data to estimate N_j for each poststratification category and ignore sampling error in these numbers. The weighted PUMS numbers are very similar to the poststratification numbers used by CBS in their raking (see Little 1996, chapter 3).

3.3 Results

We present results for four methods applied to the combined data from the seven surveys:

1. Classical estimate based on raking by demographic variables (region, sex, ethnicity, age, education, sex × ethnicity, and age × education). This is very close to the weighting method used by CBS. For estimates of results by states, we perform weighted averages within each state, using the weights obtained by the raking.
2. Regression estimate using the demographic variables and also indicators for the states, with no hierarchical model (*i.e.*, "fixed-effects" regression). This is very similar to using iterative proportional fitting to rake on states as well as demographics. The state-by-state estimates from this model should improve upon those obtained by raking on demographics because the estimates of π_j's are weighted by the population numbers N_j rather than the sample numbers n_j within each state.
3. Regression estimate using only the demographic variables, with the state effects set to zero. This model allows the average responses within states to differ only because of demographic variation; to the extent that the demographics do not explain all the variation in opinion, the model should underestimate the variability between states.
4. Regression estimate using the demographic variables, with the 48 state effects estimated with a hierarchical model (in the notation of Section 2, $L = 4$ and $K_1, K_2, K_3, K_4 = 12, 13, 12, 11$). We expect this model to perform best, both because of the flexibility of the hierarchical regression model and because the post-stratification uses the population numbers N_j.

We fit each of the regression models to the survey data, obtain posterior simulation draws for each coefficient (conditional on the estimated $\tau_1, \tau_2, \tau_3, \tau_4$), and reweight based on the PUMS data to obtain poststratified estimates for the proportion of registered voters in each state who support Bush for President.

Table 1 presents the raking estimate and the posterior medians and interquartile ranges for the three models, along with data on the survey responses and the actual election outcome. Table 2 gives the nationwide and mean absolute statewide prediction errors for the raking and the three models. The four methods give almost identical results at the national level; the real gain from the model-based estimates occurs in estimating the individual states. The reduction in mean absolute prediction error from about 6% to 5% can be attributed to using the poststratification information, with the further reduction to 3.5% attributable to the hierarchical modeling. In addition, the last two lines of Table 2 show that the uncertainty estimates from the hierarchical model are short and relatively well calibrated (slightly less than half of the true values fall inside the 50% intervals, which is reasonable since these intervals account only for sampling error and not for nonsampling errors and changes in opinion).

Figure 1 plots, by state, the actual election outcomes vs. the raking estimates and the posterior medians for the three models. As one would expect, the hierarchical model reduces variance, and thus estimation error, by shrinkage. Although the four methods correct the bias of the nation-wide estimate by about the same amount, they act differently on the individual states, with the hierarchical model performing best. Figure 2 compares the prediction errors for the hierarchical and raking estimates for the states.

Interestingly, the hierarchical model does not seem to shrink the data enough to the nationwide mean: we can tell this because, in Figure 1d, the actual election outcome is higher than predicted for low-predicted values, and lower than predicted for high-predicted values. *Under*shrinkage means that the estimated parameters $\hat{\tau}_l$ are probably *higher* than their true values, which could be caused by a pattern of nonignorable nonresponse that varies between states so that observed variability in the state proportions is caused by varying nonresponse patterns as well as actual variation in average opinions (see Little and Gelman 1996, for a discussion of this example and Krieger and Pfeffermann 1992, for a more general treatment). The undershrinkage could be quantified by comparing the estimated to the optimal level of shrinkage, but this comparison can only be made after the true values are observed.

It is also possible to compare the models by fitting each separately to each survey and examining the stability of estimates over a short period of time. This would be a more reasonable way to study the models in the common situation that the true population means never become known. Figure 3 displays, for each of our seven surveys, the estimates from raking and from the hierarchical model. (When modeling the surveys individually, we fit a common hierarchical variance for all 48 states because there was not enough data to obtain reliable maximum likelihood estimates for the four regions separately from the data in each poll.) Results are shown for the entire United States and for three representative states: California (a large state), Washington (mid-sized), and Nevada (small). For convenience, the plot also show the estimates based on the seven surveys pooled and the actual election outcomes. For all the individual states, the hierarchical estimate is less variable over time than the raking estimate. The pattern is clearest in Nevada, where the sample size for the individual surveys was so low that the raking estimate degenerated to 0 or 1 in most cases, but the better performance of the hierarchical model is clear in the other states as well. For

Table 1

By state: election results (proportion of the two-party vote in 1988 received by Bush); survey data (unweighted mean and sample size) from the combined surveys; raking estimate using CBS variables; and posterior median (and interquartile range; that is, width of the central 50% uncertainty interval) of poststratified estimates based on state effects unsmoothed, set to zero, and fit by a hierarchical model. Estimates are labelled 1, 2, 3, 4 corresponding to the descriptions in Section 3.3.

				Poststratification estimates (and IQRs)			
State	Election result	Sample size	Unweighted mean	1: Raking estimate	2: State effects unsmoothed	3: State effects set to 0	4: Hierarchical model
AL	0.60	134	0.72	0.67	0.63 (0.05)	0.56 (0.01)	0.62 (0.05)
AR	0.57	86	0.57	0.53	0.53 (0.06)	0.60 (0.01)	0.55 (0.06)
AZ	0.61	141	0.62	0.61	0.62 (0.05)	0.56 (0.02)	0.61 (0.05)
CA	0.52	1075	0.57	0.53	0.55 (0.02)	0.53 (0.01)	0.55 (0.02)
CO	0.54	126	0.59	0.59	0.58 (0.06)	0.57 (0.01)	0.57 (0.05)
CT	0.53	103	0.53	0.55	0.52 (0.06)	0.49 (0.02)	0.51 (0.06)
DE	0.56	30	0.40	0.37	0.42 (0.11)	0.60 (0.01)	0.52 (0.08)
FL	0.61	553	0.64	0.62	0.61 (0.03)	0.62 (0.01)	0.61 (0.03)
GA	0.60	211	0.62	0.58	0.56 (0.04)	0.56 (0.01)	0.56 (0.04)
IA	0.45	102	0.38	0.38	0.38 (0.06)	0.59 (0.01)	0.41 (0.06)
ID	0.63	31	0.52	0.58	0.52 (0.12)	0.59 (0.02)	0.55 (0.08)
IL	0.51	429	0.55	0.52	0.53 (0.03)	0.52 (0.01)	0.52 (0.03)
IN	0.60	215	0.75	0.73	0.74 (0.04)	0.56 (0.01)	0.72 (0.04)
KS	0.57	105	0.72	0.71	0.71 (0.06)	0.57 (0.01)	0.68 (0.05)
KY	0.56	146	0.57	0.53	0.56 (0.05)	0.64 (0.01)	0.57 (0.05)
LA	0.55	153	0.62	0.60	0.61 (0.05)	0.54 (0.01)	0.59 (0.04)
MA	0.46	277	0.47	0.41	0.46 (0.04)	0.50 (0.02)	0.47 (0.04)
MD	0.51	207	0.52	0.50	0.49 (0.04)	0.56 (0.01)	0.50 (0.04)
ME	0.56	44	0.52	0.52	0.55 (0.10)	0.52 (0.02)	0.54 (0.08)
MI	0.54	399	0.58	0.55	0.57 (0.03)	0.54 (0.01)	0.57 (0.03)
MN	0.46	210	0.54	0.53	0.53 (0.05)	0.59 (0.01)	0.53 (0.04)
MO	0.52	235	0.46	0.43	0.46 (0.04)	0.55 (0.01)	0.47 (0.04)
MS	0.61	170	0.69	0.70	0.65 (0.04)	0.53 (0.01)	0.63 (0.04)
MT	0.53	31	0.39	0.40	0.40 (0.12)	0.58 (0.02)	0.50 (0.09)
NC	0.58	239	0.59	0.60	0.55 (0.04)	0.58 (0.01)	0.55 (0.04)
ND	0.57	54	0.56	0.56	0.55 (0.09)	0.58 (0.01)	0.56 (0.08)
NE	0.61	90	0.58	0.60	0.56 (0.07)	0.58 (0.01)	0.56 (0.06)
NH	0.63	20	0.70	0.68	0.73 (0.13)	0.53 (0.02)	0.61 (0.10)
NJ	0.57	301	0.57	0.60	0.53 (0.04)	0.46 (0.01)	0.53 (0.03)
NM	0.53	87	0.55	0.54	0.57 (0.07)	0.54 (0.02)	0.56 (0.06)
NV	0.61	19	0.68	0.80	0.67 (0.13)	0.56 (0.02)	0.60 (0.09)
NY	0.48	639	0.42	0.37	0.41 (0.03)	0.45 (0.01)	0.41 (0.02)
OH	0.55	454	0.62	0.63	0.58 (0.03)	0.55 (0.01)	0.58 (0.03)
OK	0.58	93	0.57	0.62	0.59 (0.07)	0.63 (0.01)	0.60 (0.06)
OR	0.48	111	0.50	0.47	0.50 (0.06)	0.58 (0.02)	0.52 (0.06)
PA	0.51	431	0.54	0.54	0.52 (0.03)	0.48 (0.02)	0.52 (0.03)
RI	0.44	65	0.28	0.29	0.27 (0.07)	0.50 (0.02)	0.34 (0.06)
SC	0.62	151	0.70	0.67	0.66 (0.05)	0.55 (0.01)	0.64 (0.04)
SD	0.53	52	0.54	0.51	0.53 (0.09)	0.58 (0.01)	0.54 (0.08)
TN	0.58	252	0.68	0.69	0.66 (0.04)	0.60 (0.01)	0.65 (0.03)
TX	0.56	594	0.58	0.52	0.56 (0.03)	0.60 (0.01)	0.56 (0.02)
UT	0.67	61	0.80	0.85	0.79 (0.07)	0.60 (0.02)	0.72 (0.06)
VA	0.60	255	0.69	0.72	0.67 (0.04)	0.59 (0.01)	0.66 (0.03)
VT	0.52	12	0.54	0.58	0.60 (0.19)	0.53 (0.02)	0.55 (0.11)
WA	0.49	269	0.47	0.41	0.46 (0.04)	0.58 (0.01)	0.48 (0.04)
WI	0.48	264	0.49	0.53	0.48 (0.04)	0.57 (0.01)	0.49 (0.04)
WV	0.48	79	0.48	0.52	0.48 (0.07)	0.65 (0.01)	0.53 (0.06)
WY	0.61	13	0.50	0.36	0.59 (0.17)	0.59 (0.02)	0.59 (0.10)

Table 2

Summary statistics for raw mean of responses, raking estimate, and three poststratified estimates from the combined surveys. Summaries given are the estimated mean of the 48 state vote proportions weighted by state voter turnout (thus, estimated national popular vote proportion for Bush excluding Alaska, Hawaii, and the District of Columbia); the mean absolute error of the 48 state estimates; the average width of the 50% intervals for the states; and the number of the 48 states whose true values fall within the 50% intervals.

Summary	Actual result	Unweighted mean	Raking estimate	State effects unsmoothed	State effects set to 0	Hierarchical model
Mean of national popular vote	0.539	0.568	0.549	0.548	0.547	0.550
Mean absolute error of states	–	0.056	0.066	0.049	0.048	0.035
Average width of 50% intervals	–	–	–	(0.069)	(0.016)	(0.057)
Number of states contained in 50% interval	–	–	–	18	3	20

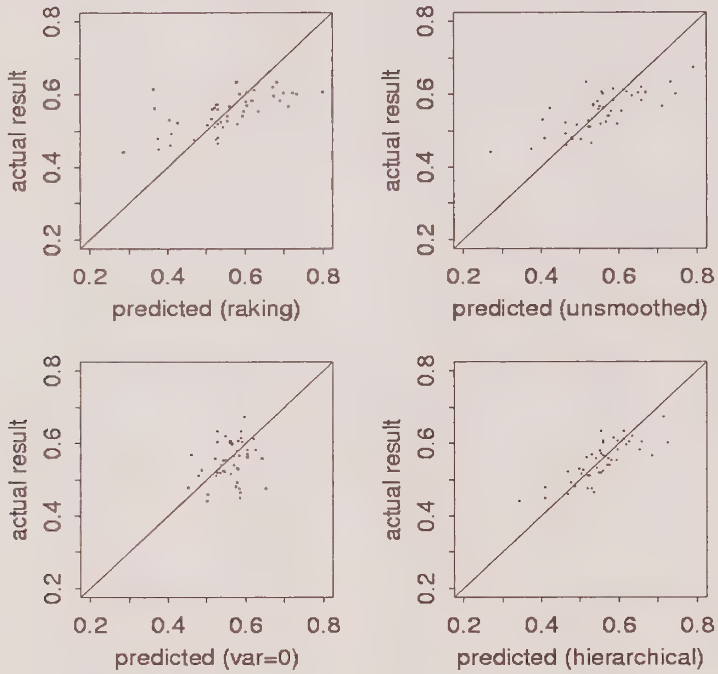


Figure 1. Election result by state *vs.* posterior median estimate for (a) raking on demographics, (b) regression model including state indicators with no hierarchical model, (c) regression model setting state effects to zero, (d) regression model with hierarchical model for state effects.

example, it was not reasonable to assign Bush only 46% of the support in California (in the poll 3 days before the election) or only 30% of the support in the state of Washington. For the United States as a whole, however, the two estimates are quite similar (in fact, when all seven polls are combined, the raking estimate performs very slightly better), indicating once again that the benefits from the modelling approach appear when studying subsets of the population.

The results for Washington have the surprising property that the regression estimate based on the combined surveys (shown at time " – 1 " on the graph) is lower than the seven estimates from the original surveys. This occurs because the data from the combined surveys show that the state of Washington supports Bush less than would be predicted merely by controlling for the demographic covariates (that prediction would be the estimate for Washington from the model with state effects set to zero, which from Table 1 is

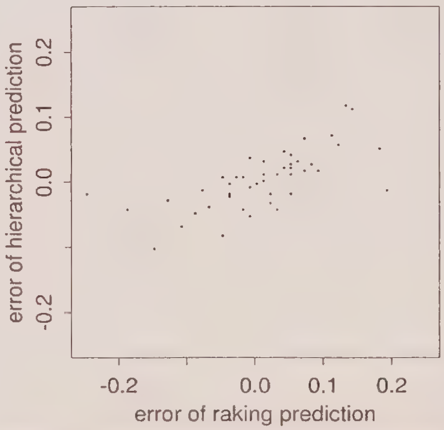


Figure 2. Scatterplot of prediction errors by state for the hierarchical model vs. the raking estimate. The errors of the hierarchical model are lower for most states.

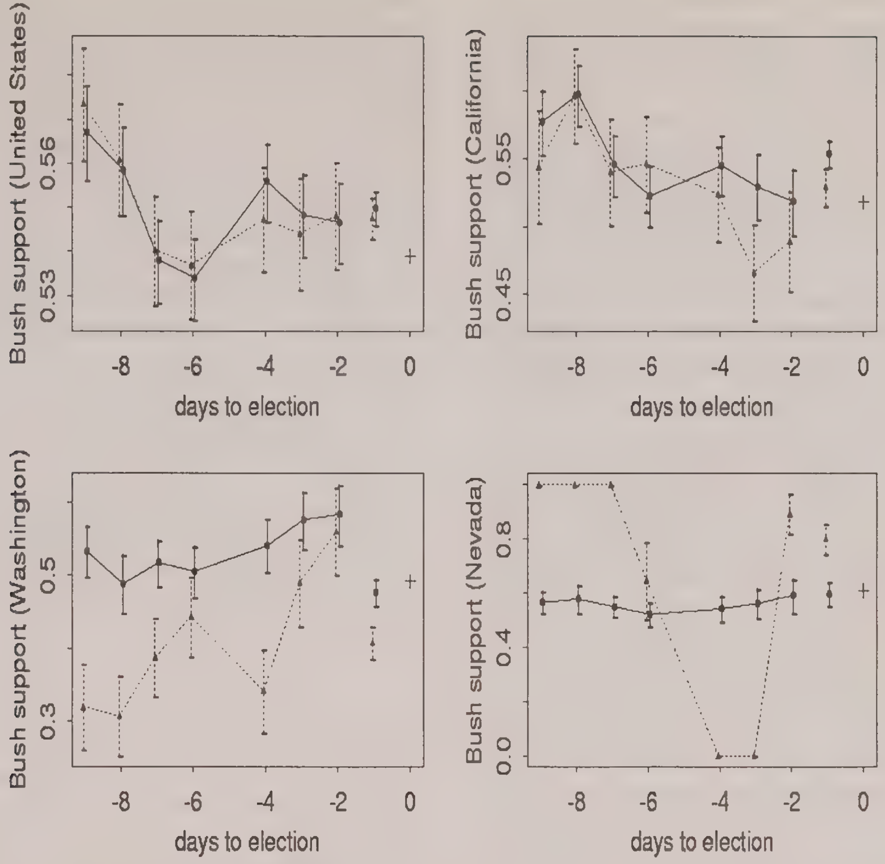


Figure 3. Estimated Bush support estimated separately from seven individual polls taken shortly before the election for (a) the entire U.S. (excluding Alaska, Hawaii, and the District of Columbia), (b) a large state (California), (c) a medium-sized state (Washington), and (d) a small state (Nevada). Each plot shows the raking estimates as a dotted line and the estimates from hierarchical model as a solid line, with error bars indicating 50% confidence bounds for the raking and 50% posterior intervals for the model-based estimates. The polls were taken between nine and two days before the election. Estimates based on the combined surveys are shown at time "−1", and the actual election result is shown at time "0" on each plot.

0.58). But none of the individual surveys, taken alone, had enough data to make a convincing case that Washington was so far from the national mean, and so the Bayes estimate shrunk their estimates to a greater extent. This behavior, while it may seems strange at first, is in fact appropriate: with a smaller survey, there is less information about the individual poststratification categories, and the model-based estimate produces an estimate for each category that is closer to the sample mean. When all seven surveys are combined, more information is available, and the model relies more strongly on the data in each category. This is how the Bayes procedure essentially balances the concerns of poststratifying on too few or too many categories.

4. DISCUSSION

Poststratification is the standard method of correcting for unequal probabilities of selection and for nonresponse in sample surveys. From the modelling perspective, raking or poststratification on a set of covariates is closely related to a regression model of responses conditional on those covariates, with population quantities estimated by summing over the known distribution of covariates in the population. Conditioning on more fully-observed covariates allows one to include more information in forming population estimates, but it is well known that raking on too large a set of covariates yields unacceptably variable inferences. We propose a method of poststratification on a large set of variables while fitting the resulting regression with a hierarchical model, thus harnessing the well-known strengths of Bayesian inference for models with large numbers of exchangeable parameters.

The Bayesian poststratification is most useful for estimation in subsets of the population (*e.g.*, individual states in the U.S. polls) for which sample sizes are small. A related area in which modeling should be effective is in combining surveys conducted by different organizations, modeling conditional on all variables that might affect nonresponse in either survey. In addition, the methods in this paper can obviously be applied to continuous responses by replacing logistic regressions by other generalized linear models.

Our purpose in Bayesian modeling is not to fit a subjectively "true" model to the data or the underlying responses, but rather to estimate with reasonable accuracy the average response conditional on a large set of fully-observed covariates. More accurate models of the responses should allow more accurate inferences – but even the simple exchangeable mixed effects model we have fit, with hyperparameters estimated from the data, should perform better than the extremes of the fixed effects model or setting coefficients to zero. Ultimately, the goal of probability modeling and Bayesian inference in a sample survey context is to allow one to make use of abundant poststratification information (*e.g.*, census data classified by sex, ethnicity, age, education, and state) to adjust a relatively small sample survey.

Difficulties with modeling approaches such as ours could arise in several ways. If one adjusts to a large number of categories using too weak a model (such as the model with unsmoothed state effects), the resulting estimates can be too variable. If the population distributions of the variables used in the poststratification are not available (for example, adjusting to a variable that is not measured or is measured inaccurately by the Census), then the N_j's must be modeled also, which requires additional work. Of course, such additional work would be required to rake on these variables as well. Since all of the methods, including raking and regression methods, assume ignorable models, they will yield incorrect inferences when unmeasured variables affect nonresponse and are correlated with the outcome of interest.

The methods described here are intended as an improvement upon raking-type poststratification adjustments and are not intended to, by themselves, correct for nonignorable nonresponse. However, by allowing one to adjust for more variables, the Bayesian poststratification should allow the use of models for which the ignorability assumption is more reasonable. Having a large number of poststratification categories (*e.g.*, in 48 states) creates problems with classical weighting methods because many categories will have few or even no respondents. Interestingly, however, having many categories can make Bayesian modeling more reliable: more categories means more random effects in the regression, which can make it easier to estimate variance components.

ACKNOWLEDGEMENTS

We thank Xiao-Li Meng and several reviewers for helpful comments and the National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824.

APPENDIX: COMPUTATION

We use an EM-type algorithm to estimate the hyperparameters τ_l; given these, we sample from the posterior distribution of the coefficients β using a normal approximation to the logistic regression likelihood. We use this approximation for its simplicity and because it is reasonable for fairly large surveys, as in our application in Section application; if desired, more exact computations can be performed using the Gibbs sampler and Metropolis algorithm (see Clayton 1996), perhaps using the algorithm described here as a starting point.

When the data distribution is normal and the means are linear in the regression coefficients, the EM algorithm can be used to obtain estimates of the variance components (Dempster, Laird, and Rubin 1977), treating the vector of coefficients β as "missing data." In this framework, the "complete-data" loglikelihood for τ_l is

$$L(\tau_l \mid \gamma_l) = \text{const} - K_l \log \tau_l - \frac{1}{2\tau_l^2} \sum_{k=1}^{K_l} \gamma_{kl}^2,$$

so the sufficient statistic for τ_l is $t(\gamma_l) = \sum_{k=1}^{K_l} \gamma_{kl}^2$. Given the current estimate τ^{old}, the expected sufficient statistic is

$$E(t(\gamma_l) \mid y, \tau^{\text{old}}) =$$

$$\| E(\gamma_l \mid y, \tau^{\text{old}}) \|^2 + \text{trace}(\text{var}(\gamma_l \mid y, \tau^{\text{old}})).$$

Since these two terms are not analytically tractable for our model, we use the following approximations which are easily obtained: (1) approximate $E(\gamma_l \mid y, \tau^{\text{old}})$ with an estimate $\hat{\gamma}_l$, based on y and the estimate τ^{old}, and (2) approximate $\text{var}(\gamma_l \mid y, \tau^{\text{old}})$ from the curvature of the log-likelihood at the estimate, $\hat{V}_{\gamma_l} = (-L''(\hat{\gamma}_l))^{-1}$. We update these approximations iteratively for all $l = 1, \ldots, L$ simultaneusly, converging to an approximate maximum likelihood estimate $(\hat{\tau}_1, \ldots, \hat{\tau}_L)$. Given an initial guess τ^{old}, the algorithm proceeds by iterating the following two steps to convergence.

Approximate *E*-step. Solve the likelihood equations iteratively, as described below. Use the estimate $\hat{\beta}$ to obtain an approximation to $E(t(\gamma_l) \mid y, \tau^{\text{old}})$, for each $l = 1, \ldots, L$.

We solve the likelihood equations $d/d\beta\, L(\beta \mid y, \tau) = 0$ using iteratively weighted least squares, involving a normal approximation to the likelihood $p(y \mid \beta) = \Pi_i p(y_i \mid \beta)$, based on locally approximating the logistic regression model by a linear regression model (see Gelman *et al.* 1995, p. 391). Let $\eta_i = (Z\beta)_i$ be the linear predictor for the i-th observation. Starting with the current guess of $\hat{\beta}$, let $\hat{\eta} = Z\hat{\beta}$. Then a Taylor series expansion to $L(y_i \mid \eta_i)$ gives $z_i \approx N(\eta_i, \sigma_i^2)$, where

$$z_i = \hat{\eta}_i + \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \left(y_i - \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} \right)$$

$$\sigma_i^2 = \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)}.$$

Let $\hat{\Sigma}_\beta$ denote the value of Σ_β based on plugging in the current estimate $\hat{\tau}$, and let $\Sigma_z = \text{diag}(\sigma_i^2)$. Then we obtain an updated estimate and variance matrix using weighted

least squares based on the normal prior distribution and the normal approximation to the logistic regression likelihood:

$$\hat{\beta} = (Z' \hat{\sum}_z^{-1} Z + \hat{\sum}_\beta^{-1})^{-1} Z' \hat{\sum}_z^{-1} z \tag{2}$$

$$\hat{V}_\beta = (Z' \hat{\sum}_z^{-1} Z + \hat{\sum}_\beta^{-1})^{-1}. \tag{3}$$

We iterate until convergence and then use $\hat{\beta}$ and the appropriate elements of $\hat{V}_\beta$ to estimate $\mathrm{var}(\gamma_l \mid y, \tau^{\mathrm{old}})$.

***M*-step**. Maximize over the parameters τ_l to obtain $\tau_l^{\mathrm{new}} = (\hat{E}(t(\gamma_l) \mid y, \tau^{\mathrm{old}})/K_l)^{1/2}$, for each $l = 1, ..., L$. Set τ^{old} to τ^{new} and return to the approximate E-step.

Once the approximate EM algorithm has converged to an estimate $\hat{\tau}$, we draw β from a normal approximation to the conditional posterior distribution $p(\beta \mid y, \hat{\tau})$, using the values from equations (2) and(3) at the last EM step as the mean and variance matrix in the normal approximation. For each draw of the vector parameter β, we compute the category means, $\pi = \mathrm{logit}^{-1}(X\beta)$, and any population totals of interest, counting each category j as N_j units in the population.

REFERENCES

BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., and ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *Journal of the American Statistical Association*, 88, 1149-1166.

CLAYTON, D.G. (1996). Generalized linear mixed models. In *Practical Markov Chain Monte Carlo*. (Eds. W. Gilks, S. Richardson, and D. Spiegelhalter), 275-301. New York: Chapman & Hall.

DEMING, W., and STEPHAN, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals of Mathematical Statistics*, 11, 427-444.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.

GELMAN, A., CARLIN, J.B., STERN, H.S., and RUBIN, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

GELMAN, A., and KING, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23, 409-451.

HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society*, 142, 33-46.

KRIEGER, A.M., and PFEFFERMANN, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, 18, 225-239.

LAZZERONI, L.C., and LITTLE, R.J.A. (1997). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, to appear.

LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics,* 7, 405-424.

LITTLE, R.J.A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association,* 88, 1001-1012.

LITTLE, T.C. (1996). Models for nonresponse adjustment in sample surveys. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

LITTLE, T.C., and GELMAN, A. (1996). A model for differential nonresponse in sample surveys. Technical report.

LONGFORD, N.T. (1996). Small-area estimation using adjustment by covariates. *Qüestió,* 20, to appear.

NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics,* 5, 223-239.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika,* 63, 581-592.

VOSS, D.S., GELMAN, A., and KING, G. (1995). Pre-election survey methodology: details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59, 98-132.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 137-146
Statistics Canada

Estimating the Population and Characteristics of Health Facilities and Client Populations Using a Linked Multi-Stage Sample Survey Design

K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN and G. NARAYANA[1]

ABSTRACT

This paper demonstrates the utility of a multi-stage sample survey design that obtains a total count of health facilities and of the potential client population in an area. The design has been used for a state-level survey conducted in mid-1995 in Uttar Pradesh, India. The design involves a multi-stage, areal cluster sample, wherein the primary sampling unit is either an urban block or rural village. All health service delivery points, either self-standing facilities or distribution agents, in or formally assigned to the primary sampling unit are mapped, listed, and selected. A systematic sample of households is selected, and all resident females meeting predetermined eligibility criteria are interviewed. Sample weights for facilities and individuals are applied. For facilities, the weights are adjusted for multiplicity of secondary sampling units served by selected facilities. For individuals, the weights are adjusted for survey response levels. The survey estimate of the total number of government facilities compares well against the total published counts. Similarly the female client population estimated in the survey compares well with the total enumerated in the 1991 census.

KEY WORDS: Sample survey; Program evaluation; Health services; Developing country.

1. INTRODUCTION

The evaluation of the impact of health programs on population-level health outcomes often requires knowledge of the number and characteristics of facilities and potential clients. Such information is frequently lacking in developing countries where program record keeping and vital registration systems tend to be incomplete and poorly maintained.

To obtain current information on health status, health service use, service performance, and client needs, programs have resorted to occasional sample surveys, often designed and conducted independently and subareally (Aday 1991; Ross and McNamara 1983). Some demographic and health surveys (Macro International 1996), however, do provide a national profile of population-level health outcomes, such as fertility, child mortality, and nutritional well-being. The distinct advantage of a national population sample for planning health programs is its ability to measure the attitudes and behaviors of clients as well as non-clients. Program service statistics are limited to actual clients and may not yield the most current or accurate picture of service use.

In addition to client behaviours, it is useful to monitor the accessibility and quality of services, but this requires a separate review of service provision at health facilities or related outlets. Efforts in developing countries, like the situation analysis studies (Miller, Ndhiovu, Gachara and Fisher 1991), involve probability surveys of health facilities and can provide a national overview of program performance. However, often they are restricted to reviewing public health programs because of incomplete registration of private health providers, such as private clinics or pharmacies. The lack of complete and accurate registration of private-sector service providers prevents probability sample surveys from being used to monitor health care patterns through this sector.

Constraints on available resources to expand and improve the delivery of health care in developing, as well as developed, countries are increasing. This suggests that a more efficient use of resources available for monitoring and evaluation, particularly through surveys, is a consideration for all concerned. Innovative approaches to sample surveys should be developed to provide health planners and managers with a maximum of information at a minimum of precision loss.

We present results from a multi-stage, cluster sample survey designed to estimate the population and characteristics of health facilities and target client populations. The cluster sample for the survey, conducted in the large northern Indian state of Uttar Pradesh, is used as a basis for selecting health facilities and households, with subsequent selection of service staff from the facilities and of married women of childbearing age from the households. The survey was designed to generate independent samples of health facilities, staff, households, and client populations for the health services.

The next section of this paper will describe the survey design, its contents, and fieldwork procedures as applied in

[1] Kaushalendra K. Singh, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Statistics, Faculty of Science, Banaras Hindu University, Varanasi 221005 India; Amy O. Tsui, Director, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Maternal and Child Health, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Chirayath M. Suchindran, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Gaade Narayana, The Futures Group International, 1050 17th Street, N.W., Suite 1000, Washington, DC 20036.

Uttar Pradesh. The following section presents the comparative results on health facilities and population, and the last section will discuss lessons learned for survey design from the Uttar Pradesh application. These lessons will be important specifically for this survey's planned replication in two years but generally informative for other countries that may adopt the linked design.

2. THE PERFORM SURVEY IN UTTAR PRADESH

The PERFORM (Project Evaluation Review For Organizational Resource Management) Survey was designed to measure benchmark indicators for a large family planning project called the Innovations in Family Planning Services (IFPS) project sited in Uttar Pradesh and co-funded by the Government of India and the U.S. Agency for International Development. Uttar Pradesh has a population of over 140 million and by itself would rank as the fifth largest developing country.

2.1 Content

Indicator estimates for IFPS are needed at three levels: (1) public and private service delivery points (SDPs), (2) service providers staffing the SDPs or facilities, and (3) client population, represented by women of reproductive age. As IFPS seeks to improve the family planning service environment, it is imperative to obtain measures of indicators at this level but in such a way as to be relatable to the women resident in those environments.

As a result, the PERFORM survey developed seven questionnaires:

1-2) An urban block and village questionnaire to inventory all potential and actual providers of health services in the sampled village or urban block;

3) A fixed service delivery point (FSDP) questionnaire to gather information on the staff, services, equipment, supplies, and education and motivation activities at sampled public and private facilities.

4) A staff questionnaire administered to all FSDP staff involved in family planning services (identified from the FSDP questionnaire) to assess their capabilities and service experiences;

5) An individual service agent (ISA) questionnaire to all individuals working outside of self-standing facilities (FSDPs) who currently or potentially can provide health planning services, such as private doctors, pharmacists, midwives, lay health workers, and retailers;

6) A household questionnaire to be administered to heads of the sampled households to enumerate household members and selected demographic and social characteristics;

7) An individual questionnaire for currently married women between the ages of 13 to 49 (identified from the household questionnaire) to collect information on knowledge of and past, current, and intended use of health services, recent pregnancy and contraceptive behaviors, and additional background characteristics.

2.2 Sampling Design

PERFORM was designed to provide estimates of facility and population characteristics at the state, regional, divisional, and district levels. The district was important since it was the focal point for introducing innovative approaches and additional IFPS inputs. At the time of the survey design, Uttar Pradesh had 14 administrative divisions; two districts were selected from each using probability proportional to size (PPS) procedures. These areal units have administrative-political boundaries and thus public administration utility. The districts were also aggregated into five regional groupings.

In each district, the total number of households to be sampled was fixed at 1,500. A sample of 1,500 households per district was determined to be sufficient to provide estimates for the main population level indicators. An overall target sample size of 1,627 ever-married women aged 13-49 was required to detect a change of 5 per cent point in contraceptive prevalence (with $\alpha = 0.05$ and $1 - \beta = 0.90$) at district level. It is expected that the number of ever-married women aged 13-49 per household would be 1.15 and therefore, by visiting a sample of 1,415 households the required number of ever-married women would be obtained. Allowing for an increase of 5 per cent to accommodate non-response and non-availability, a target sample of 1,725 ever-married women aged 13-49 from the 1,500 households was considered to be sufficient. The schematic diagram of the sample design is given in Figure 1.

The districts were further stratified into rural and urban areas. According to the Census of India, all places with a municipality, a municipal corporation, a cantonment board, a notified area committee, or all other places with a minimum of 5,000 population, with at least 75 percent of the male working population engaged in non-agricultural pursuits and a population density of at least 400 persons per square kilometer, are classified as urban areas. Urban blocks and rural villages served as the secondary sampling units (SSUs). The 1,500 households to be sampled from each district were allocated to the rural and urban areas in proportion to the size of population within the district. However, if the allocated proportion of urban population was less than 20 percent, the allocation of households in the urban area was fixed at 20 percent. This allocation was prescribed to ensure coverage of a sufficient number of health delivery points.

Households within rural areas were selected using a stratified two-stage sampling plan. The villages in the rural areas were first stratified into four strata depending on the size of the of the population as follows:

Stratum	Population size of the village
I	100 - 499
II	500 - 1,999
III	2,000 - 4,999
IV	5,000 and above.

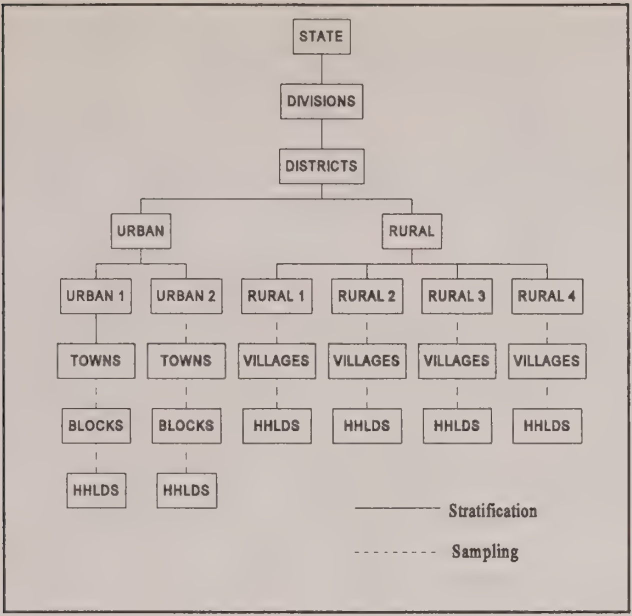


Figure 1. Schematic Diagram of PERFORM Sample Design

Villages with fewer than 100 residents or 20 households were excluded from the list (such villages were rare in the present study). The number of villages to be selected from each district was allocated proportionally to each of the four strata. Villages were selected by first arranging them within the stratum by the female literacy rates and then selecting the required number of villages by a PPS sampling procedure. All households in the selected villages were listed and mapped, and a target number of 20 households was drawn from each selected village using systematic sampling. Villages with more than 500 households or with a population size of 2,500 or more (some in stratum III and all in stratum IV) were segmented into four parts, and two segments were selected for household listing and selection. The required 20 households were selected taking ten households from each segment using systematic random sampling.

Households in urban areas were also selected using a stratified two-stage sampling plan. The towns in the urban areas of a district were stratified into two strata according to population size as follows:

Stratum	Population size of the town
I	100,000 and more
II	Fewer than 100,000.

All towns within stratum I were selected with certainty. Towns in stratum II were arranged according to population size and the required number of towns were selected by PPS. From each sampled town a minimum of two blocks were selected using PPS methods. All households in the selected blocks were listed and mapped, and 15 households were selected from each urban block using systematic random sampling.

2.2.1 District Selection Probability

Let m_k denote the population of the k-th district within a division. Because two districts must be selected from each division, the probability of selecting the k-th district from a division r_k is obtained as

$$r_k = 2 * \frac{m_k}{M}$$

where M is the total population of the division ($M = \sum_{k=1}^{t} m_k$) and t is the total number of districts in the division.

2.2.2 Village and Household Selection Probability

Let n_{ijk} denote the number of households in the i-th village, j-th stratum and k-th district. Then, p_{ijk}, the probability of selecting village i from the j-th stratum and k-th district is obtained as,

$$p_{ijk} = a_{jk} * \frac{n_{ijk}}{N_{jk}} * r_k$$

where a_{jk} and N_{jk} are, respectively, the number of villages selected and the total number of households in the j-th stratum and k-th district.

Let q_{ijk} be the probability of selecting a household from the rural areas of a selected district. Then q_{ijk} may be given as

$$q_{ijk} = p_{ijk} * \frac{20}{n_{ijk}}$$

where 20 is the number of households drawn from the selected village.

The weights for villages and households are then the inverse of their selection probabilities, *i.e.*, $1/p_{ijk}$ and $1/q_{ijk}$, and are denoted as VW_{1ijk} and HW_{1ijk} respectively.

2.2.3 Town, Urban Block and Household Selection Probability

The probability of selecting the j-th town from the k-th district, t_{jk}, is obtained as

$$t_{jk} = 1 \qquad \text{if the population of the town is} > 100{,}000$$

$$t_{jk} = c_k \frac{s_{jk}}{S_k} \qquad \text{if the population of the town is} < 100{,}000$$

where s_{jk} is the total number of households in the j-th town (with a population < 100,000) in the k-th district, c_k is the number of towns selected in district k, and S_k is the total number of households in towns with less than 100,000 population in district k.

Let u_{ijk} denote the probability of selecting the i-th urban block from the j-th town and k-th district. Then u_{ijk} is obtained as

$$u_{ijk} = b_{jk} * \frac{x_{ijk}}{Y_{jk}} * t_{jk} * r_k$$

where b_{jk} is the number of urban blocks selected and Y_{jk} is the total number of households in the j-th town and k-th district, and x_{ijk} is the number of households in the i-th block, j-th town and k-th district.

The probability of selecting a household from the i-th urban block and the k-th district, denoted as v_{ijk}, is given as,

$$v_{ijk} = u_{ijk} * \frac{15}{x_{ijk}}$$

where 15 is the number of households drawn from the selected urban block.

The weights for urban blocks and households are then the inverse of their selection probabilities, *i.e.*, $1/u_{ijk}$ and $1/v_{ijk}$, and are denoted as UW_{1ijk} and HW_{1ijk} respectively. Since the population-level estimates are based on individuals, all individuals in a selected household received the household weight. No selection procedure was used for eligible respondents within a household.

2.2.4 Adjustment for Household Questionnaire for Non-response and Over-sampling of Urban Blocks

The adjustment of the household weight for non-response is done under the assumption of random non-response within the village (or urban block) and is carried out as follows:

Let n_1 be the number of households selected and n_2 be the number of households where interviews are completed. Then the adjusted weight for households due to non-response is defined as

$$HW_{2ijk} = HW_{1ijk} * \frac{n_1}{n_2}.$$

The final household weight also includes an adjustment of proportion of urban population in the district, where an over-sampling of urban blocks has occurred (districts with less than 20 percent of urban population).

Let n_3 be the actual proportion of urban population in a district and n_4 the proportion of urban population in the sample. Then the adjusted weight for households due to non-response and over-sampling of urban blocks is defined as

$$HW_{3ijk} = HW_{2ijk} * \frac{n_3}{n_4}.$$

2.2.5 Selection of Service Delivery Points in Sample Districts

To obtain a probability sample of service delivery points, FSDPs and ISAs were selected in relation to the SSUs, *i.e.*, the villages or urban blocks, as follows:

1) All private and public sector health institutions in selected rural and urban SSUs;
2) All sub-centres, primary health centres, community health centres, post-partum centers providing services to the population in the selected rural SSUs;
3) All private hospitals with 10 or more beds in the nearest town (with fewer than 100,000 population) within 30 kms of selected rural SSUs;
4) All municipal hospitals, district hospitals, and medical college hospitals;
5) All clinics and hospitals runs by voluntary agencies, the organized sector, and cooperatives; and
6) All ISAs in selected villages and urban blocks.

It is probably helpful first to describe the organized delivery of health care through the government sector. Residents of all villages are entitled to obtain health care from a government sub-centre (SC), a primary health centre (PHC), and a community health centre (CHC). Villages with 5,500 population or more often have an SC located within their boundaries. Approximately six SCs will report to one PHC, and PHCs in turn are linked to a CHC. At times the PHC is integrated with the CHC; as a result, our estimation must be of CHCs and PHCs combined, while SCs are estimated separately. (Population growth has led to the establishment of "additional PHCs" and redistricting of the original PHC catchment areas. These additional PHCs have been included in the estimation of the number of PHCs.) All SCs assigned to a sampled village were visited, as were their affiliated PHCs and CHCs.

At the time of listing and mapping households in each urban block and village, the FSDPs and ISAs were also listed and mapped. In addition, key informants in each SSU were interviewed regarding health outlets not visibly obvious. The selection of service delivery points – FSDPs and ISAs – within the SSU boundaries, or affiliated with the government's health subcentre, involved a full census. The one exception to this was for municipal hospitals, district hospitals and medical colleges, which were self-selected and thus had a weight of unity. The selection probabilities of the other FSDPs and ISAs are then a function of the probability of selecting the SSU, and the inverse of the latter serves as the weight of the FSDP or ISA unit. Weights for CHCs, PHCs, and SCs were calculated with the procedure below after determining some fieldwork "failure" in selecting these types of facilities correctly. (This failure is discussed later.)

Since CHCs and PHCs are associated with more than one SSU, we have assumed that one PHC exists per 30,000 population (which is approximately the actual average for Uttar Pradesh) and that one SC serves approximately 5,500 (actual district averages range from 4,000 to 6,500). Under this assumption, the CHC/PHC weight for each selected SSU is then

$$W_{CHC/PHC} = \frac{\text{Total population in selected SSU}}{30{,}000} * VW_{1ijk} \text{ (or } UW_{1ijk})$$

and the SC weight for each selected SSU is

$$W_{SC} = \frac{\text{Total population in selected SSU}}{5{,}500} * VW_{1ijk} \text{ (or } UW_{1ijk}).$$

All weights for FSDPs that were not self-selected had to be adjusted for multiplicity, *i.e.*, when an FSDP was selected into the sample on the basis of more than one SSU. For example, a CHC/PHC might be selected because of two sampled SSUs. In this case, the weight for the CHC/PHC was the sum of the weights of the two selected SSU, *i.e.*, $W_{\text{CHC/PHC}}$, associated with its selection.

2.3 Survey Implementation

Fieldwork for the PERFORM Survey was conducted from June to September 1995 in Uttar Pradesh. The survey was executed by four organizations contracted following a competitive procurement process. One organization that had tested the PERFORM survey design in one district a year earlier served as the nodal or coordinating organization. Master training to survey project coordinators and supervisors was provided, including a field pretest. The actual fieldwork for PERFORM was carried out in six-member teams composed of 1 male supervisor, 1 female editor, 1 male interviewer and 4 female interviewers. Each fieldwork organization on average engaged 3 teams to cover one district, or a total of 18 field staff for data collection per district (or 21 teams for a total of 126 field staff to cover 7 districts). Overall field supervision was the responsibility of a specially–appointed four–member team, one assigned to each consulting fieldwork organization. Following field editing, the questionnaires were transported to the home offices of the survey organizations for data entry and cleaning. One type of staff person, the auxiliary nurse–midwife who is stationed at a subcentre, was difficult to reach, even after the standard three attempts.

3. RESULTS

Table 1 gives the sample coverage for the PERFORM survey, in terms of the number of units selected of each type, the number successfully interviewed, and the completion rate. The completion rates are very high for ample units requiring personal contact – ranging from 94.3 for eligible women to 96.7 percent for households. Interview completion rates were 95 percent for facilities and agents. Only for fixed facility staff was the rate somewhat lower at 90 percent, a respectable although not an outstanding level. (One type of staff person, the auxiliary nurse-midwife who is stationed at a subcentre, was difficult to reach, even after the standard three attempts.)

3.1 Population Size and Characteristics

We compare first population–level measures on selected demographic indicators obtained from other sources with those from the PERFORM survey, as shown in Table 2. The figures indicate that PERFORM results compare favorably with census measures as well as these from the recent National Family Health Survey (NFHS) conducted in Uttar Pradesh in late 1992 and early 1993, with a sample size of 11,438 ever-married women aged 13 to 49. The enumerated population shows a growth of almost 10.5 million persons since the 1991 census, and the percentage of households in urban areas is close across all three sources. The ratio of women to men is slightly lower in PERFORM (891) than in the NFHS (917). The percentage of the population in the two age groups (0 to 14 and 65 and over) compares well, as does the percentage of households belonging to the scheduled castes. The percentage of households belonging to scheduled tribes is 3.1, higher than the 1.1 observed in the NFHS. This may reflect an actual growth in such households with increased in-migration to large towns and cities by scheduled tribe members. The proportions literate show small gains since the NFHS but compare well overall. The total fertility rate and the level of modern contraceptive use also are similar and change in a consistent direction between the dates of the two Uttar Pradesh surveys. Results in Table 2 suggest that PERFORM's sample design, based on traditional multistage cluster sample designs used for demographic surveys, was executed properly to produce state-level results comparable to the census and earlier NFHS survey. The standard error and design effect of the estimates were also given in the Table

Table 1
Coverage of Sample Units of PERFORM Survey: Uttar Pradesh, 1995

Sample Coverage	Sample Units						
	Villages	Urban Blocks	Households	Eligible Women	Fixed SDPs	FSDP Staff	Individual Agents
Number Sampled	1,539	738	42,006	48,009	2,549	7,026	23,364
Number Interviewed	1,539	738	40,633	45,277	2,428	6,320	22,335
Percent completed	100.00	100.00	96.7	94.3	95.3	89.9	95.6

Notes: Villages and urban blocks served as the primary sampling units; eligibility criteria for women were currently married and between ages 13 to 49 years; SDP = service delivery point.

Table 2
Basic Demographic Indicators for Uttar Pradesh, India

Index	Uttar Pradesh				
	Census (1991)	NFHS (1992-93)	PERFORM (1995)	Standard Error	Design Effect
Population	139,112,287	*u*	149,758,641	1,542,952	–
Percent urban	19.8	22.6[a]	21.6[a]	0.6553	12.6095
Sex ratio[b]	879	917	891	34.1010	0.9727
Percentage 0-14 years old	39.1	41.8	40.2	0.1306	1.9049
Percent 65+ years old	3.8	4.8	4.7	0.0513	1.5789
Percentage scheduled	21.0	18.0[a]	20.0[a]	0.3790	3.6536
Percentage scheduled tribe	0.2	1.1[a]	3.1[a]	0.1818	4.4694
Percent Literate[c]					
Male	55.7	65.3	67.6	0.3352	6.4634
Female	25.3	31.4	37.4	0.3824	8.6821
Total	41.6	49.9	53.3	0.3352	12.2385
Total fertility rate	5.1	4.8	4.5	–	–
Modern contraceptive	*u*	18.5[d]	22.0[d]	0.3499	3.4111

u = Unavailable
[a] Based on number of households
[b] Number of females per thousand males
[c] Based on population aged 7 and above for the census and population aged 6 and above for NFHS and PERFORM
[d] Percentage of currently married women aged 15 to 49 using modern contraceptive method.

In Table 3 we compare the age and sex distributions for Uttar Pradesh obtained from the NFHS and PERFORM, as well as from the Sample Registration System, operated by the Office of the Registrar General. The sex ratios for the two surveys are also given. The age-sex distributions are again comparable across the three sources. However, there is a markedly lower sex ratio for the age group 30-49 years (820) in PERFORM and a slightly higher one for ages 50-64 (993) than those in the NFHS (941 and 960 respectively). We suspect some of this difference is due to a "push" of females out of the end of childbearing ages by field investigators of *one* survey organization to avoid completion of the pregnancy calendar and history portions of the questionnaire. (Upon further investigation, we found the sex ratios for women aged 50-64 to be uniformly higher in the seven districts under one organization's responsibility than those of others.) As a result, there are somewhat more women aged 50-64 enumerated in the PERFORM Survey than may actually be the case. This also may mean that births to women who were actually under age 50 were under-enumerated. Because this is not a high-fertility age group, the bias is not likely to be large.

3.2 Facility Size and Characteristics

By visiting and interviewing the facilities selected through the SSUs or cluster, we are able to generate an independent sample of health facilities and service providers. (These include those who currently, as well as potentially can, provide family planning services, *i.e.*, not all the estimated number of retail outlets (general merchant, kirana and pan shops) shown presently dispense contraceptives.) The weighted counts of these outlets is shown in Table 4. Our ability to validate the estimates of independent agents is weakened by the fact that many of them are not registered, particularly the "unqualified" (or quack) doctors. Narayana, Cross and Brown (1994: Table 8) report a 1991 total number of 112,568 villages in Uttar Pradesh, which would suggest almost one traditional birth attendant per village and 1 anganwandi worker for every 4.5 villages on average. These ratios appear reasonable given known circumstances regarding access to such types of care. The figures are quite close and provide evidence of the utility of the linked cluster sample design.

Table 3
Percent Distribution of the De Jure Population by Age and Sex, Based on SRS, NFHS, and PERFORM Sources for 1991-95

Age	SRS (1991)		NFHS (1992-93)			PERFORM (1995)		
	Male	Female	Male	Female	Sex Ratio	Male	Female	Sex Ratio
0-4	14.4	14.4	14.6	14.6	917	13.8	14	909
5-14	24.9	24.4	27.5	26.0	868	27.2	26.3	861
15-29	28.4	26.8	25.1	26.4	967	25.4	27.7	972
30-49	20.7	21.9	19.2	19.7	941	19.8	18.3	820
50-64	8.2	8.5	8.4	8.8	960	8.6	9.6	993
65+	3.6	4.0	5.2	4.4	718	5.2	4.1	702
Total	100.0	100.0	100.0	100.0		100.0	100.0	

Source for sample Registration System (SRS): Office of the Registrar India (1993a)
Source for NFHS: National Family Health Survey, Uttar Pradesh (1992-93)

Table 4
Total Number of Estimated Public and Private Sector Delivery Points by Type in Uttar Pradesh, India: 1995

Fixed service delivery points	Number	Individual service agents	Number
Total	31,400	Total	1,099,825
Hospitals		Physicians	
Government allopathic	968	Private resident allopathic	32,182
Government ISM	688	Private visiting allopathic	9,011
Municipal allopathic	57	Private resident (unqualified)	62,880
Municipal ISM	23	Private resident ISM	42,343
Private	5,212	Private visiting ISM	9,138
Private voluntary	130	Anganwadi workers	25,994
Private ISM	35	Village health workers	65,532
Industrial	61	Traditional birth attendants	110,546
Medical colleges	9	Medical shops	40,979
CHC/PHC/Additional PHC	3,948	General merchants	133,517
Subcentres	20,151	Kirana shops	376,679
Other	137	Pan shops	136,353
		Depot holders	5,818
		Other	48,855

3.3 Estimation Approaches

The estimated number of CHC/PHCs and SCs in Table 4 is based on the assumption that each such facility serves a fixed population size, *i.e.*, 30,000 and 5,500 respectively – the figures used by the government for planning health service delivery. The precision of the estimation would have been improved if the actual size of the local catchment population were known. In the absence of this information, we have used a constant population estimate for these two facility types.

Alternate estimation approaches were used prior to arriving at the above procedure. The first is illustrated in Table 5, which presents the actual and weighted counts of CHC/PHCs and SCs in each of the 28 survey districts. These figures are based on weighting the selected facilities by the SSU size only and without adjusting for multiplicity. The PERFORM sample selected in a total of 633 CHC/PHCs or 34.8 percent of the total (1818) and 1,267 subcenters or 13.3 percent to the total (9,491) in the 28 districts. These can be compared against the actual numbers

of CHC/PHCs and SCs in 1995 obtained from the Uttar Pradesh Department of Health and Family Welfare. It is evident that this weighting approach substantially over-estimates the number of CHC/PHCs (3,472 compared to 1,818) but yields a nearly identical number of SCs (9,495 compared to 9,491). Using the villages and urban blocks as SSUs is reasonable as they are the public administration units (and population sizes) used to determine the location of subcenters.

They, however, do not offer an adequate stratification basis for the larger health facilities. Precision is lost because we weight with the inverse of the SSU's population and when CHC/PHCs are selected in for very small SSUs, the associated weight is disproportionately inflated. This results in a higher-than-actual count of such facilities, a situation most problematic in two districts – Allahabad and Sultanpur. If these two districts are eliminated, the over-estimation is 22.5 (± 0.8) percent instead of 91 percent. (Under-estimation of CHC/PHCs results where the reverse occurs, as in Bareilly district. Because of PPS, large stratum IV villages have small weights, and in fact most selected FSDPs in this district have been sampled in the SSUs of this size.)

A second estimation approach used was to calculate the expected number of CHC/PHCs and SCs based on *a priori* knowledge that such facilities were located in SSUs of minimum size 30,000 or 5,500, respectively. With 1991 census information on the SSU population, we reconstructed the distribution of each district's population by stratum size and divided each stratum by the CHC/PHC or SC catchment size (30,000 or 5,500 respectively). This provides the expected number of CHC/PHCs and SCs for each district. We can compare this with the observed number of such facilities, obtained at the time of fieldwork where local community informants were asked whether there was a CHC/PHC and/or SC located within the SSU. This comparison is shown in Table 6, which also includes a fieldwork organization code (I to IV) in the event any pattern of survey error is evident. This approach overestimates the number of subcenters by 19.6 percent and under-estimates the number of CHC/PHCs by 26.5 percent. Excluding the two districts with a high number of stratum I SSUs (Allahabad and Sultanpur) reduces the CHC/PHC underestimation to 10.2 percent. Tabulation of estimation bias by fieldwork organization shows no systematic bias.

The results from the two weighting approaches suggest that the SSU offers an appropriate measure of size (MoS) for the selection of subcenters, since its average population size may approximate the SC's catchment size of 5,500. A larger MoS may have served the selection of CHC/PHCs better, since this facility's catchment size covers those for five to six subcenters. Because SSU size is the basis for the weight for CHC/PHCs, when the selected SSU is small, the bias in estimated counts can be large. A future design to consider is to use a cluster of SSUs that are contiguous to the selected SSU and have an MoS similar to the catchment size of CHC/PHCs. The probability of such a facility being present within the boundaries of the SSU cluster will then be higher and the weight, constructed on the basis of the total population in the SSU cluster, more reliable. In other words, our estimation is limited by not knowing how many SSUs are served by one CHC/PHC.

Table 5

Total Actual and Estimated Total Number of Community Health Centres, Primary Health Centers,[a] and Subcentres by District in Uttar Pradesh, India: 1995

District	CHC/PHC		Sub-centre	
	Actual	Estimated	Actual	Estimated
Aligarh	77	69	399	369
Azamgarh	103	69	475	949
Almora	44	104	254	468
Allahabad	112	981	594	677
Ballia	73	93	357	485
Banda	89	101	322	302
Bareilly	71	42	355	162
Dehradun	24	41	139	60
Etawah	69	84	323	364
Fatehpur	57	73	309	327
Firozabad	33	34	234	236
Gonda	107	183	528	461
Gorakhpur	59	84	470	460
Jhansi	51	77	251	157
Kanpur Nagar	12	13	81	74
Maharajgang	30	39	195	180
Meerut	76	187	410	119
Mirzapur	64	69	309	302
Moradabad	92	81	485	248
Nainital	53	79	287	344
Rampur	37	19	170	139
Saharanpur	60	49	293	388
Shahjahanpur	52	59	301	298
Sultanpur	70	487	394	649
Tehri Garhwal	31	5	159	63
Unnao	63	162	344	106
Sitapur	87	44	437	450
Varanasi	122	144	616	658
Total	1818	3472(±21)	9491	9495(±15)
Total[b]	1636	2004(±13)		

[a] Includes additional primary health centres
[b] Excludes Allahabad and Sultanpur districts
Source for 1995 actual figures from Government of Uttar Pradesh Department of Medical and Family Welfare.

Table 6
Observed and Expected Sampled Number of CHCs/PHC[a] and Subcentres Within the Rural Village (Urban Block) by District in Uttar Pradesh, India: 1995

District	CHC/PHC		Sub-Centre		Field Work Company
	Actual	Estimated	Actual	Estimated	
Aligarh	6	5	10	17	II
Azamgarh	3	5	24	15	III
Almora	5	2	14	9	I
Allahabad	19	4	17	18	III
Ballia	9	7	34	27	III
Banda	8	9	19	27	III
Bareilly	5	3	10	16	II
Dehradun	5	7	10	21	I
Etawah	8	7	17	20	II
Fatehpur	9	7	22	25	IV
Firozabad	6	6	28	30	II
Gonda	8	5	15	18	IV
Gorakhpur	5	4	16	20	IV
Jhansi	7	6	16	24	II
Kanpur Nagar	2	2	6	8	II
Maharajgang	4	4	9	13	IV
Meerut	12	8	12	34	II
Mirzapur	7	7	22	22	III
Moradabad	5	5	9	19	I
Nainital	6	4	19	19	I
Rampur	2	5	14	16	I
Saharanpur	6	6	25	21	I
Shahjahanpur	5	3	14	15	II
Sultanpur	16	6	21	15	IV
Tehri Garhwal	1	3	3	10	I
Unnao	3	6	17	17	IV
Sitapur	10	6	9	24	IV
Varanasi	6	5	18	18	III
Total	186	147	450	538	
Total[b]	151	137			

[a] Includes additional primary health centres
[b] Excludes Allahabad and Sultanpur districts.

4. DISCUSSION

The cluster-based sample design for generating independent samples of facilities and households, which can be analyzed individually or jointly, does warrant more extensive consideration in data collection efforts for health program research and evaluation in developing countries. Careful design and fieldwork sampling and execution can yield high-quality and acceptably precise survey estimates, as our results show. The weighted totals, rather than sample totals, themselves are numbers useful to program planners who decide the flow of personnel, material, and financial resources to and among various facility sites and area locations. The linkage of facility to individual records offers further important analytic opportunities to assess the relative importance of personal background and service supply factors on health outcomes of interest (*e.g.*, Boyd and Iversion 1979).

At the same time, our application of this design reveals several lessons. First there is an obvious need to monitor the survey fieldwork closely with increased on-site data entry so that the apparent "push" of eligible women out of the older age ranges can be prevented. This is difficult to detect through individual questionnaire spot checks but can

be observed in aggregate tabulations produced, say, weekly on completed questionnaires. Second, the excess count of CHCs/PHCs in two districts, where the survey fieldwork involved two *different* organizations suggests that stratum I villages might have been disproportionately selected or that some of the CHCs/PHCs reported to be within the SSU boundaries were in fact not. The former may have occurred as a sampling error since each fieldwork organization was provided with a list of sampled SSUs. Third, the listing and mapping of SSUs for facilities, individual health care providers, and households are an important stage of the fieldwork. Careful execution of this task allows the sampled units to be re-located for future follow-up. This will be an essential measurement effort for evaluating the IFPS project.

Certainly for a survey as complex as PERFORM, scaled to capture the levels of and differentials in the patterns of health service delivery and client use in an area as populous as Uttar Pradesh, the fact that the quality of the data meets most standards of precision evidences an important fieldwork achievement as well as design innovation.

ACKNOWLEDGMENTS

Partial support for this study has been provided by The EVALUATION Project, USAID Contract #DPE-3060-C-00-1054-00. The views contained herein are solely those of the authors and not the sponsoring agency. The authors acknowledge with appreciation earlier assistance on the sample design from Daniel Horowitz and T.K. Roy. We thank Lynn Moody Igoe of Carolina Population Center for editing the paper. Authors are also thankful to the anonymous referees for their useful comments and suggestions.

REFERENCES

ADAY, L.A. (1991). *Designing and Conducting Health Surveys: A Comprehensive*. San Francisco: Jossey-Bass Publishers.

BOYD, L.H., Jr., and IVERSION, G.R. (1979). *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.

MACRO INTERNATIONAL, INC. (1996). *Demographic and Health Surveys Newsletter*, 8, 1-12.

MILLER, R.A., NDHIOVU, L., GACHARA, M.M., and FISHER, A.A. (1991). The situation analysis study of the family planning program in Kenya. *Studies in Family Planning*, 22, 131-143.

NARAYANA, G., CROSS, H.E., and BROWN, J.W. (1994). Family planning programs in Uttar Pradesh issues for strategy development: tables. Centre for Population and Development Studies, Hyderabad, India.

ROSS, J.A., and McNAMARA, R. (Eds.) (1983). *Survey Analysis for the Guidance of Family Planning Programs*. Liege, Belgium: Ordina Editions.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 147-156
Statistics Canada

Computer-assisted Interviewing in a Decentralised Environment: The Case of Household Surveys at Statistics Canada

J. DUFOUR, R. KAUSHAL and S. MICHAUD[1]

ABSTRACT

In 1993, Statistics Canada implemented Computer-assisted Interviewing (CAI) for conducting interviews for some household surveys that were conducted in a decentralised environment. The technology has been successfully used for a number of years, and most household surveys have now been converted to this collection mode. This paper is a summary of the experience and the lessons that have been learned since the research started. It describes some of the tests that led to the implementation of the technology, and some of the new opportunities that have arisen with its implementation. It also discusses some challenges that were faced when CAI was implemented (some are on-going issues), and ends with a brief overview of where this may lead us in the future.

KEY WORDS: Household surveys; Data collection; Computer-assisted interviewing; Decentralised environment.

1. INTRODUCTION

The first systems of computer-assisted interviewing (CAI) were developed in the early 1970s (see Nicholls and Groves 1986). These systems were mainly developed by market research organisations in the United States and, a little later, independently by well-known university research centres. During the late 1970s and early 1980s, computer-assisted interviewing systems became much more sophisticated, and their use expanded greatly. By the late 1980s, a number of universities and survey research centres in the United States had a computerised collection system (see Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin 1997). Clark, Martin and Bates (1997) provide an overview of the development and implementation of such systems in four major government statistical agencies.

In 1987, Statistics Canada conducted its first experiment with computer-assisted interviewing for household surveys. At that time, the tests were done in a "centralised telephone collection environment". The series of tests with computer-assisted interviewing was extended into the early 1990s to try to adapt to the more general collection methodology.

At Statistics Canada most household surveys share a common sampling frame and data collection environment. The main user of this frame is the monthly Labour Force Survey (LFS). Data collection is decentralised with the initial interview in person at the selected dwelling and the subsequent five interviews by telephone from the interviewer's home. To accomplish this, almost a thousand interviewers have been equipped with portable computers. Interviewers are attached to one of the five regional offices located throughout Canada. A number of household surveys in the bureau follow a similar collection strategy by subsampling from the Labour Force Survey sample, by administering a series of supplementary questions after the Labour Force Survey interview or by contacting persons who have formerly participated in the survey. As a result, not only is the Labour Force Survey sample shared with other surveys, but so is the collection infrastructure. All interviewers are required to work on the Labour Force Survey for a specified week each month, and for the rest of the time, they have been trained and equipped to collect data for other surveys. For further details on the Labour Force Survey methodology, see Statistics Canada (1998).

The 1990s saw testing of the implementation of the computer-assisted collection mode not only for the LFS but also for other surveys sharing that common infrastructure and having very different requirements. The results of the various tests led to the implementation of computer-assisted interviewing for the LFS in November 1993 (Dufour, Kaushal, Clark and Bench 1995) while its supplementary monthly surveys have been changed gradually. In January 1994, a new longitudinal survey, the Survey of Labour and Income Dynamics (SLID) was launched using computer-assisted interviewing (see Lavigne and Michaud 1995). Since then, the National Population Health Survey (NPHS) along with the National Longitudinal Survey of Children and Youth, (NLSCY) introduced in August and November 1994 respectively, have also adopted this collection mode (see Tambay and Catlin 1995, Brodeur, Montigny and Bérard 1995). For further details on the structure and implementation of this computerised collection mode in longitudinal surveys, see Brown, Hale and Michaud (1997). Today most of Statistics Canada's household surveys are collected using a computerised mode and a common infrastructure.

This article focuses primarily on methodology aspects of decentralised computer-assisted interviewing for household surveys. We provide an overview of the implementation process for the statistical agency as a whole, a brief discussion of the challenges associated with the new collection vehicle and a list of references for more detailed information on specific topics. Despite "growing pains", Statistics Canada is continuing to experiment with and

[1] J. Dufour and R. Kaushal, Household Survey Methods Division; S. Michaud, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

implement this new technology in various surveys to render these surveys more cost efficient and to improve data quality and the survey monitoring process.

The article is divided into five sections. In the next section, aspects of implementation are discussed with reference to several surveys. Section 3 details new opportunities arising from computer-assisted interviewing. The ongoing challenges and new problems that surveys face as a result of using a decentralised computerised collection mode, as well as the changes that are taking place, are discussed in Section 4. The last section describes the future of CAI for household surveys at Statistics Canada.

2. FIRST YEARS OF IMPLEMENTATION

Adopting a computerised collection method for household surveys held the promise of several benefits: (i) a decrease in survey costs, (ii) better data quality, (iii) the possibility of using more complex questionnaires, (iv) data made available more quickly, (v) a tool for tracing operations, (vi) the possibility of using dependent interviews, and (vii) a generalised collection method for all of the agency's household surveys. However, these benefits were not realised overnight, or without effort. Ongoing evaluations and adjustments were required in the introduction and stabilisation phases.

Despite a number of tests being conducted before the implementation of CAI, unforeseeable problems occurred with the adoption of this method, but over time, they became less frequent and easier to solve. In addition, during this period, the series of quality indicators analysed carefully by different groups of Statistic Canada experts were somewhat disrupted. It took about one year to realise the anticipated benefits. This section describes the main points in the process of changing from the traditional paper approach to computer-assisted interviewing, where collection and capture are integrated.

2.1 Centralised Computer-assisted Telephone Interviewing

The traditional approach to interviewing used a paper questionnaire filled out in pencil to facilitate edits made by the interviewer. Often such an approach is referred to as Paper and Pencil Interviewing (PAPI). In this traditional mode, an interviewer edited the questionnaire to ensure that the information was correct and complete. Information abbreviated to shorten the interview was filled-in in detail after the interview and before the form was sent for data capture. The first change towards computerisation was the use of Computer-assisted Telephone Interviewing (CATI). This computerised collection mode was used for surveys that were conducted by telephone from a central location. CATI was the first instance of amalgamation of the collection and capture of information in household surveys. Given the state of technology at that point, the computers capable of handling the complexity associated with computer-assisted interviewing were fairly large. Hence, CATI could replace PAPI only in centralised telephone surveys. In the 1990s, with the advent of more powerful portable computers decentralised CAI replaced PAPI. A decentralised collection mode is, in effect, what is used in most household surveys. In addition, data collection often required the ability to do either telephone interviews or personal visits. However, much of the know-how and experience of computer-assisted telephone interviewing could be applied to decentralised computer-assisted interviewing.

Since the 1980s, it was the Labour Force Survey (LFS) that served as the main research and testing vehicle for CATI technology. The first test, conducted in 1987, was a controlled study that compared CATI in a centralised environment to PAPI. It consisted of a research project carried out jointly between Statistics Canada and the US Bureau of the Census (see Catlin and Ingram 1988). The study showed that there were differences between the two collection methods in terms of data quality indicators, and those differences were in favour of CAI in terms of lower rejection rates on edit, reduction in path errors on the questionnaire and decrease in undercoverage in the LFS.

While CATI was never implemented for the LFS, the experience was used to set up a CATI facility for use in random digit dialling (RDD) in household surveys. As technology progressed, CATI was used to collect more complicated RDD surveys like the General Social Survey (GSS) and the Violence against Women Survey. Computer-assisted telephone interviewing continues to be used as an integral part of household collection at Statistics Canada complemented by the computer-assisted interviewing infrastructure.

2.2 Technological Testing

A new wave of testing began in the early 1990s as part of the decennial redesign of the LFS (Singh, Gambino and Laniel 1993; Drew, Gambino, Akyeampong and Williams 1991). The launching of three large scale longitudinal surveys by Statistics Canada made the investment for a CAI infrastructure possible by sharing the costs among a number of surveys. Consequently, in 1991, a second test was conducted using the LFS and SLID to study the feasibility of using new technologies (see Williams and Spaull 1992). Portable computers which require the use of a stylus rather than a keyboard for entering data were tested. The results showed that the technology was promising but that it needed further improvements for it to be used to handle the requirements of Statistics Canada's household surveys.

The following year, from July 1992 to January 1993, a third and a fourth test were conducted, this time using conventional portable computers. The results for the LFS are documented in Kaushal and Laniel (1995), while the results for SLID are reported in Michaud, Le Petit, and Lavigne (1993) and Michaud, Lavigne and Pottle (1993). For the LFS, the main objective of this third test was to determine if the transition to the new technology would disrupt the LFS data series. The secondary objective of the test was to determine whether the new technology affected

data quality and interview costs. Additional objectives of this test were the operational development and evaluation of the CAI approach. For the longitudinal surveys, the main concern was the length and complexity of the questionnaires and the addition of new functions, such as tracing. Consequently, the main criterion in assessing the application was the feasibility of developing various functions. The results showed that CAI had no major impact for the LFS on either the data series disseminated, the survey's main quality indicators, or interview costs. On the strength of general comparisons with outside sources and an analysis of missing variables, the new technology was adopted.

2.3 New Dimension of Nonresponse

With the adoption of CAI, there was an unintentional development of a new dimension of nonresponse that is due to "technical problems". Such nonresponse resulted from cases that were lost or not received before the end of the collection period. The PAPI version of this type of nonresponse was related to occasional postal problems. Conceptually, these situations do not refer to real nonrespondents; however, the information is not available in time to produce estimates.

These technical problems assume three different forms: (i) transmission problems, (ii) equipment problems, and (iii) unavoidable problems. Transmission problems are the most common. They arise, for example, when telephone lines are down, when there is a problem with the automatic downloading of data, when an attempt is made to download data while maintenance is being carried out on the mainframe computer, or simply because of a malfunction in the CAI system. The second type of problem, although less common, occurs when a hard drive crashes, the magnetic tape drive fails, there is insufficient memory or there are computer equipment problems at the regional offices. Finally, unavoidable problems, which are even less common, include specific problems implicitly created by the above two categories, for example when only one of the two components expected from a respondent is transmitted or if the initialisation parameters needed for the proper functioning of the programs are missing.

Nonresponse due to technical problems diminished over the initial months. This component of nonresponse was analysed quite carefully to explain an upward trend in nonresponse and to assess the performance of the CAI approach (see Simard, Dufour and Mayda 1995; Dufour, Simard and Mayda 1995). At the start of the conversion of the household surveys to CAI, technical problems represented on average 15% of total nonresponse and could alone explain up to 25% of nonresponse. It took almost a full year before any significant reduction was observed in this component of nonresponse. Today, in 1997, the nonresponse due to technical problems is practically non-existent.

In the first year, the bulk of the problems were due to a conflict over memory management in the notebook computer between two pieces of software used in case management. This was resolved by a re-write of a part of the software, which eliminated the conflict and made the system more efficient. The more subtle issues of the transition were communication and experience. A communication strategy was developed to enable the different players (in particular technical personnel and interviewers) to better understand each other, disseminate information more quickly and adequately inform all persons concerned. When CAI was first introduced, it took technical support personnel more than a day to find a solution to some problems. Faster response procedures were established, and a 24-hour support service was set up at head office in Ottawa. With such a substantial change, a learning and adjustment period is required, and Statistics Canada was no exception.

2.4 Impact of CAI on Nonresponse

Are there grounds for believing that the use of CAI had an effect on nonresponse rates? The answer to such a question has to be yes in light of the technical problems encountered, primarily at the beginning of the conversion process. However, if this aspect of the nonresponse is discounted, there is no indication that CAI had any lasting effect on nonresponse rates. The LFS nonresponse fluctuated following the introduction of CAI, but these fluctuations may be explained by a number of other factors (the redesign of the sample, which is now more urbanised; hiring of new interviewers; *etc.*), since the LFS was undergoing a major overhaul. It took just under two years for overall nonresponse to return to levels similar to those recorded in the paper and pencil era.

In the LFS, the conversion took place over a period of five months during which time the CAI and PAPI nonresponse rates could be compared. These comparisons show that the nonresponse rates for CAI (excluding technical problems) and those for PAPI were in the same range and exhibited the same trends (see Simard and Dufour 1995). Moreover, all the main components of nonresponse, namely refusal to participate in the survey, household temporarily absent, no one at home and other reasons, exhibited similar annual patterns before and after the implementation. There were concerns that respondents would be more reluctant to answer due to the presence of a computer for personal interviews, resulting in an increase in refusals. However, no change in the refusal component was detected.

In early 1995, the three longitudinal surveys (SLID, NLSCY and NPHS), as well as the LFS, were conducted during similar collection periods. The current case management environment, as well as the sharing of the infrastructure among surveys, created extra pressure on interviewers in the field. Moreover, the survey collection periods were limited because there was a limited number of applications that could reside on the computers at the same time. Analysis was done to determine if response problems arose from conducting several surveys simultaneously, or in quick succession, in the field using CAI. For the quarterly collection of the NPHS, interviewers followed-up nonrespondents in previous collections. An analysis was

carried out to determine the possible conversion rate. The results showed that in the case where there were fewer CAI surveys in the field at the same time, a first wave of follow-ups of nonrespondents increased the response rate, but continuing the process for a second or third time brought few gains (an increase of 5.76% from the first to the second quarter, 0.97% from the second to the third, and 0.91% from the third to the fourth). However, a last follow-up was carried out in June 1995 when there were almost no surveys in the field. This procedure improved the overall response rate by approximately 5%, which was higher than expected. This led to the conclusion that CAI had to be able to give more flexibility in the length of the collection period and allow multiple applications to reside on the computer in order to maintain the response rates that would have been obtained in a paper and pencil environment.

3. NEW OPPORTUNITIES FOR HOUSEHOLD SURVEYS

The adoption of CAI collection has added new opportunities to household surveys. These new opportunities, which were either non-existent or operationally difficult in a paper and pencil mode, help to reduce non-sampling errors, to collect more specialised information, to facilitate the reconstruction of family units and to make contact with family units that break apart or merge. In fact, this collection method is better suited to adjust the collection process according to the changing needs of today's society.

3.1 Dependent Interviews

The introduction of the new technology served to resolve household survey problems that had proven intractable under the traditional paper and pencil interview approach. In particular, CAI helped to increase the information that could be provided by the interviewer to a respondent contacted for the second time for the reduction of (i) response error (coding, capture or recall error), in particular the seam problem and telescoping, and (ii) response burden by confirming the information instead of requesting it again (or by requesting only partial information).

The seam problem has been documented for longitudinal surveys in Murray, Michaud, Egan and Lemaître (1990), which notes that the problem arises in reconciling data from successive collection periods. If no reconciliation has been attempted between collections, an artificially large change in estimates is generally observed at each collection transition. This problem is generally explained by respondents' difficulty in pinpointing the date when a change occurs. As to telescoping, it results from a tendency to include certain events that occurred outside the reference period.

Under the traditional PAPI approach, the type of information that could be provided to interviewers was limited. Questionnaires could only be pre-printed with basic information, as there were physical limits to the amount of information that could be pre-printed, especially for long questionnaires. In some cases, additional information was even printed on a separate questionnaire. This procedure also involved additional logistical problems for the interviewer. The use of information from earlier occasions in the process is known as feedback. With computer-assisted interviewing, feedback is made possible in two ways: proactively and reactively. A discussion of this is also provided in Brown *et al.* (1997).

Proactive use of feedback is used to reduce response error by helping the respondent to situate him/herself. For example, SLID gathers detailed information on a maximum of six jobs in the previous year. Without feedback, the name of the employer or the occupation might be written slightly differently, and a job that continued over a period of two years could be incorrectly classified as a change. Initially there was some concern that the respondent would perceive feedback negatively, but in fact, few negative comments have been received.

The confirmation rate is generally high – over 90% for data that are presented to the respondent (see Hale and Michaud 1995). The study of Hiemstra, Lavigne and Webber (1993) concerning the labour market suggests that while feedback generally serves to reduce the seam effect, the problem is only partially solved. For example, SLID confirms employment, job search or joblessness at the beginning of the previous calendar year over a one-year recall period. Micro-comparisons with a cross-sectional monthly survey, conducted over the first five months of the year, suggest that feedback greatly reduces the seam effect. However, consistency with cross-sectional data decreases over the months, which seems to suggest that response error, although eased by feedback, is still a problem.

The proactive use of feedback may, however, underestimate measures of change. For this reason, for sensitive information and for reasons of confidentiality, the technique is also used reactively. The reactive use of feedback can be used to detect unusual changes, or to confirm inconsistencies in the data. As an illustration, in the interview for the first wave of SLID, jobless spells are identified and for each spell the respondent is asked whether employment insurance benefits have been received. The second wave interview asks for detailed information on various sources of income and amounts received including employment insurance benefits. Comparisons with outside sources suggest that traditionally, the amounts of employment insurance reported in a survey represent approximately 80% of the contributions paid. In SLID, previous information was stored in memory. If an amount was not reported and there was an indicator flagging an inconsistency with the first-wave interview, an additional question was asked to determine whether the amount had been omitted. An analysis of the first wave of SLID suggests that reactive checking increased the number of reported cases by nearly 30%. However, 28% of these persons who had neglected to report an amount, confirmed that they had received an amount but were unwilling to

report that amount. There was thus confirmation of the source, but the amount had to be imputed and the problem was not totally solved. More details on this subject may be found in Dibbs, Hale, Loverock and Michaud (1995).

3.2 A More Efficient Tool

With an efficient collection tool like CAI, it is now possible to collect, to limit, to access and to transfer detailed information which would traditionally have been very difficult, or even not possible, to do with PAPI.

3.2.1 Matrix of Relationships Between the Various Members of a Household

Household surveys create different levels for analysis such as the economic family and the census family, by using the relationships between the various persons in the household with a single person often called the "family head". There are limitations to this method for example, in identifying the children of blended families or reconstructing families to three generations. In a longitudinal context, the concept of family head is a definition that can vary over time and so a number of longitudinal surveys have used a matrix of relationships for all members of the household. CAI can limit collection to the lower diagonal of the matrix. Provided that the composition of a household does not change between two collections, it is not necessary to re-ask it for the relationship matrix. Interactive edits (based on age, for example) serve to correct any relationships captured in reverse (*e.g.*, a parent-child relationship). It took a number of attempts to develop an effective means of identifying relationships that would allow not only for the collection of the information but also for easy correction. With the improved version of the collection procedure, less than 1% of relationships required further correction after collection (as compared to 5.3% inconsistency before the interactive edits on the relationship matrix). Corrections in a CAI environment probably continue to be one of the areas in which research is still required.

3.2.2 Access to More Sophisticated Collection Instruments

CAI has also provided access to more sophisticated collection instruments. For example, the NLSCY obtains a variety of information on a cohort of children aged 0-11 years. One part of the interview is designed to measure the child's vocabulary level. The survey uses the Peabody Picture Vocabulary Test (PPVT) as one of its collection instruments. However, the PPVT is normally used in a more specialised environment, and persons administering it generally need several days of in-depth training since the test involves a series of images, and the child is asked to choose the image that corresponds to a given word. The starting level depends on the child's age. Questions are administered until the child gets a certain number of wrong answers. At this point, the interviewer must return to the starting level and re-administer the previous questions, until the child gives a pre-determined number of wrong answers. The administration of the test calls for determining a threshold based on criteria, counting the number of wrong answers, skipping between questions depending on the number of wrong answers, and stopping the test. These procedures would have required a considerable amount of training if it had been necessary to administer the test on paper. CAI has greatly facilitated the process by allowing programming of the edit rules in advance. The data from the first collection suggest that the computer-assisted conditions of administration yield good-quality results when compared to external norms.

3.2.3 Establishing Longitudinal Links

In the case of longitudinal links, it may happen that all the members of an initial household may be part of the longitudinal sample, as in SLID for example. In subsequent collections, the longitudinal persons are interviewed along with all persons with whom they live. In the case of a household that splits, a new household must be created for the persons who left the original household. With the adoption of CAI, it became possible to create new unique household identifiers linked to the original identifiers, this made it easier to reconcile the dynamics of change in household composition. A particular problem that has been greatly lessened is the treatment of the real duplicates that occur as a result of changes in household composition. For example, an adolescent might belong to a given household at the time of the first collection, then leave his parent's household by the time of the second collection but return to the original household by the time of the third collection. In the second collection, the person is identified as belonging to a new household, and a new identifier is thus associated with him. In the third collection, when the parents' household is again contacted, the adolescent who has returned may be indicated as a new person in the household. If the interviewer is shown the list of persons who have formerly been part of the household, the need to reconcile duplicates is greatly reduced. A similar treatment has been carried out for jobs where a list of previous employers is used for longitudinal reconciliation of jobs.

3.2.4 Tracing of Individuals

With the conversion to CAI, certain procedures such as tracing were automated. Brown *et al.* (1997) gives specific examples. As noted above with respect to establishing longitudinal links, traced individuals may all be put into a new household with a unique identifier. Fewer paper manipulations are required, and it is now possible to obtain more management information. CAI has made it possible to set up a two-level tracing procedure. The interviewer first attempts the tracing. If this is not successful, all information on the case is transferred to a tracing unit in the regional office where more sources for tracing are available. Automation has eliminated many manipulations and transcriptions of records on paper. Formerly when a household split, a new identification sheet was usually created on paper with a link to the previous household. The

names of the persons who had moved were entered on it. If the person to be traced was not found, all the forms for all the persons who had been living together the previous year were transferred. These manipulations greatly increased the risk of error. Transfers of cases between tracing levels are also done more quickly. In addition, each call is recorded automatically along with its result. While there was a similar procedure with the paper and pencil approach, the information was seldom entered. It was also hard to analyse the information for determining the most useful tracing sources.

Tracing is a key factor in maintaining data quality. With current tracing procedures, cases requiring tracing can be kept in the field a little longer, but the collection window remains limited. It is possible that more effective procedures can be established if the efforts of the various longitudinal surveys are integrated. Increased functionality, combined with central tracing, is currently being examined. This would make it possible to combine the tracing efforts of the various surveys, and it might also make it possible to have batch entries to try to link cases requiring tracing to databases.

3.3 New Quality Indicators

The CAI approach adopted by Statistics Canada for its household surveys features a complex system capable of monitoring survey activities during the collection period to ensure their smooth operation. This system called the "case management system" (CMS), is a sophisticated system that manages all survey activities from the beginning to the end of the survey cycle. This system is flexible, since it can be adapted to the requirements of the different household surveys that use it. The CMS performs three main functions: (i) routing of cases, (ii) reporting of activities and (iii) assisting interviewers. The routing component directs the movements of cases during the survey, whether from an interviewer to the regional office, from the regional office to head office, *etc*. The second component of the CMS produces different reports for describing the status of the survey at a given point in time, evaluating the performance and progress of the survey, and describing the status of interviews. A whole range of information is generated by this second component of the CMS. Lastly, the third module enables interviewers to perform their tasks more effectively, by giving options for making appointments, recording notes and so on.

As a result, this system provides a mass of information on what is actually happening in the field during a survey; every action taken on a case is recorded by the CMS. The main challenge with such a system is to avoid getting lost in the great mass of information available. Work teams have been set up to master these information sources, develop new quality indicators using this information or combining it with information already available, find uses (*e.g.*, additional training, improvement of the collection instrument), and develop ways to present these indicators effectively.

A large number of quality indicators have been produced (see Simard *et al.* 1995; Allard, Brisebois, Dufour and Simard 1996) on a regular basis at different levels of interest (geographic, interviewers, administrative). These indicators may be grouped into two categories: informational and for monitoring purposes. Examples of informational indicators are: number of attempts before completing a case, distribution of interviews completed per day of collection, best day-hour combination for reaching a respondent, median duration of interviews, and number of edit rules triggered and ignored or triggered and acted upon (see Brisebois, Dufour, Lévesque 1997). Information indicators are used to improve or make changes to the collection strategy or process.

In terms of monitoring, a series of indicators are used to trace irregularities, technical or human, in the field. Among these are: calls and visits done after the date of transmission but before the survey week, calls and visits done after Sunday of survey week, working period too early, working period too late, interviews too short, *etc*. This information serves to show whether instructions issued by head office are followed, and whether some interviewers require additional training. However, all data need to be analysed with caution to determine the cause of the irregularity. For example, an interview conducted at 4:30 am may well be at the request of a respondent, like a farmer, or due to an incorrect time on the computer clock (see Brisebois *et al.* 1997).

CAI also offers interviewers the opportunity to include a comment for each question or to explain the reason for the code used. It is therefore possible to develop adequate training, to better understand the surveys and accordingly to adapt them to realities in the field. For example, this feature made it possible to conduct a special study on the reasons for refusal to participate in one of Statistics Canada's household surveys; to conduct such a study would have formerly required a great deal of effort (see Allard, Dufour, Simard and Bastien 1996).

4. ONGOING CHALLENGES OF CAI

This section describes long-term challenges in developing, implementing and understanding the use of CAI for survey applications. The powerful tools provided by CAI have led us to degrees of complexity in content, software and electronic communications that may not be widely appreciated. The conversion to CAI has implied a new dependence on informatics. This dependence is one of the major challenges that Statistics Canada has to face with CAI, since the technology is changing so quickly.

4.1 Workload of Interviewers

A common infrastructure requires the sharing of limited resources, such as trained interviewers equipped with portable computers, by different surveys. As a consequence, any increase in either the number of surveys or the amount of information collected must be carried out jointly with the

other surveys. It should be noted that the same interviewers tend to be used by a large number of surveys, which can result in fairly large workloads, exacerbated by a short collection period. While response rates have recovered since the introduction of CAI, a heavy workload for interviewers can lead to deterioration in data quality, owing to fewer follow-ups and higher nonresponse.

Given the nature of the CMS, an administrative structure for communication, based on the needs of a given survey (based on the response codes), must be put in place to provide for the routing of cases between the interviewers, their supervisors and the regional offices. Since CAI was first introduced, there have been great improvements in the communications process to ensure that all interviewers correctly receive their assignments, the latest version of the application or various changes; nevertheless, this process must be constantly monitored. For example, after the end of the collection period, cases must be transmitted and deleted from the interviewers' computers. Often, the cases that were not transmitted consist mainly of nonresponse cases. The fact that these cases are not transmitted to head office after the end of collection means that the reasons for nonresponse are sometimes lost. While many of these problems can be detected during testing, the fact remains that a few exceptional cases still remain.

4.2 Control Procedures for CAI

The CMS and survey applications have the potential to generate many databases. The quantity of data is often overwhelming, and the data are not currently being used to their maximum potential. In addition, the speed inherent in CAI sometimes does not allow for sufficient time and resources to analyse and control this mass of information. For the moment, this information is used after the fact, but it would be highly desirable to be able to use it while the survey is in the field.

This information should be made available to interviewers in an integrated format. However, a balance is needed to avoid excessive surveillance where interviewers focus more on the quality indicators than on the quality of the data. Ideally, analysis across several surveys could identify specific problems, which could then be dealt with in training kits that are brief and focused. In addition, response rates and coverage rates could be integrated for surveys. All this information could be used to achieve more efficient time management or to develop training in specific interview skills.

4.3 Editing During Collection

While CAI offers the possibility of including a great number of edit rules at the time of the interview, it is important here as well to maintain a balance between the rules programmed into the collection instrument and the rules applied during batch processing at head office. The rules programmed into the instrument prolong the interview, which results in an increase in both costs and response burden. Over time, and with rapid changes in technology, it should be possible to apply a larger number of edits during the interview without interfering with its flow. On the other hand, clarifications at the time of the interview undeniably result in better quality data. The NPHS obtains better quality data in the second quarter by using information from the first quarter to feed the edit system. For example, clarifying with the respondent at the interview, led to the discovery that, for the arthritis variable, of the 7.0% of individuals who indicated a change in condition between the two quarters, 3.3% actually experienced a change while 3.5% represented errors. For further details, see Catlin, Roberts and Ingram (1996).

With CAI, it is also possible to store information to identify which edit rules have been triggered and what corrections were made. A study of the most frequently triggered edit rules would determine which rules most affect data quality, with the results of these studies serving not only as information but also as inputs, for changing overly strict edit rules and also for sustaining a dynamic correction system. Another aspect that is just as important is the ease with which the interviewer can make the necessary corrections. If the corrections can be made to the actual response or the preceding response to a question, the interviewer can easily identify the changes to be made. If the correction involves editing between several answers, then the need to determine which one requires correction, and to move between the various answers in which there may be an error, sometimes makes the process too complex for the edit to be carried out during the interview.

Apart from technical problems, there are methodological problems associated with the effect of edit rules on data quality. At what stage are the different edit rules the most effective? The rules that affect the flow of the questionnaire and those that determine which persons are outside the scope of the survey, are critical edit rules. The key variables used for poststratification and key estimates are best resolved at the time of the interview. The quantity of edit rules that can be incorporated into the CAI system must be balanced with the speed of the portable computer. In addition, when some edit rules are being developed for the instrument and others for central processing, care must be taken to ensure that the two types of rules are not contradictory.

4.5 Data Confidentiality

Maintaining data confidentiality, as stipulated by the *Statistics Act*, is one of the fundamental requirements of the use of CAI and the systems that support it. To meet such a requirement, a number of procedures have been developed including a computing environment with two communication networks, one external and the other internal. The data are transferred physically, by tape, from the external network to the confidential internal network since there is no link between these two networks. It is impossible to access the internal network using a public modem. Confidentiality is also ensured by encryption of data whenever they must be transmitted over telephone lines. In addition, an access control system is incorporated into all portable computers, enabling only the interviewer to access

the information. The data are also encrypted while residing on the notebook.

The challenges relating to confidentiality in a CAI environment are quite different from those encountered with PAPI. Dependent interviews offer such a challenge for SLID. Information available from the preceding wave family unit may become sensitive in the case of, say, a family break-up. Thus, while the new technology offers the benefits of dependent interviews, these are accompanied by drawbacks that must be analysed for the specific situation.

With the arrival of audio-CASI (known by the acronym CASI-A), sensitive subjects may be handled more easily. With this interview technique, respondents are linked to the computer with earphones, and the questions are read by a digitised voice. Since the question is heard via the headset, the respondent can choose whether or not to display the question on the screen. With these features, the respondent can complete the questionnaire in total anonymity. The NLSCY is planning to begin using this collection instrument by the year 2000.

4.6 Re-Interview Programs

CAI offers some enhancements over PAPI-based re-interview programs. Firstly, the rapid electronic transmission of data reduces discrepancies due to recall and memory problems since re-interview can be conducted quicker after the initial interview. Strict adherence to reconciliation procedures built into the software provides more accurate estimates of measurement error. This would eradicate the problem of interviewers peeking at the questionnaire before starting the re-interview. As well, reconciliation can be done after a subset of questions, a section or at the end of the questionnaire and as many times as desired. Re-interview cases are easily automated and integrated into a quality control process based on characteristics of the interviewer or the interview (*e.g.*, specific cases related to training issues, cases belonging to a specific group, *etc.*). The quality of the data is better since a great number of edit rules, identical to the ones used during the interview, are programmed for the re-interview. The features available from the CMS are also an asset for the re-interview program: progress of the re-interview program, performance and progress of the re-interview, easy transfer of cases, *etc*.

4.7 Interviewer Training

With the adoption of CAI, interviewers had to cope with a major change in their work method. Training was therefore an essential stage in enabling them to adapt effectively to the computerised collection method. They became familiar with new work tools, including the keyboard, the portable computer and all the computer procedures, such as saving data, charging batteries and transmitting by modem. They also had to adapt their interview style to the requirements of CAI. New interviewers, for their part, had to familiarise themselves with survey concepts, interview techniques and the collection instrument. To meet this challenge, Statistics Canada developed a training strategy based on the experience acquired during the previous testing, as well as on the experience of British and American colleagues.

Interviewer training will always be one of the key factors in the success of Statistics Canada surveys, and the agency is continually innovating in this field. For example, one of the initiatives for the LFS is a training strategy to enable senior interviewers to regularly receive a small CAI assignment (approximately 15 cases), just so they can practice collection by this method and thereby stay abreast of changes in the CAI application. In addition to the regular practice cases that are always available on the computer, the CAI system will provide interviewers with modules integrated into the collection system, dealing with such complex subjects as coverage and multiple dwellings, to enable them to always be updated or to review various difficult concepts.

5. FUTURE OF CAI AT STATISTICS CANADA

In the new environment of limited resources and high response burden, collection is becoming increasingly customised. While business surveys have been doing it for some time, mixed collection is beginning to be in demand for household surveys. Centralised collection outside the collection window for a limited number of respondents can be used to improve response rates (to focus on tracing for example). The environment necessary for this type of collection more closely resembles a CATI environment in which shared database functions for a small sample are available, with call planning functions.

A complete redesign of the CAI application and the case management system is expected to be completed by the turn of the century. In this redesign, work teams must take account not only of computer capacity but also of the human factor. The latter factor is important since data collection and data quality depend on it. Interviewers must read the screen and enter the responses, tasks that call for perceptual and motor skills different from those required for pencil and paper interviews. The wording of questions is also harder to read on the screen, and interviewers mention that it is now harder to visualise the overall structure of a questionnaire. Hence special attention must be paid to screen design, the choice of colours, the amount of text displayed, the key functions pre-programmed and the ease of moving between screens. Since interviewers are also asked to work on several surveys, an effort should be made to standardise screen formats as much as possible.

As regards the hardware and software components, work teams are currently concentrating on choosing the best combination. At present, different softwares are used for different components of some surveys. In order to standardise the applications available as much as possible, there are plans to use a uniform platform for all surveys in a Windows environment. The Windows environment should give both interviewers and programmers greater

flexibility. The security systems must also be redesigned to conform to the technology adopted and to satisfy the requirements of Statistics Canada. Harmonisation of questions among surveys should be attempted, which would allow CAI programming to become more modularised. Respondent burden would also be reduced.

The new system will have to be able to take account of both past and present requirements. For example, system features are re-examined in the light of the progress reports provided to operational staff in order to determine which areas need improvement. As noted in Section 4, a number of other possibilities are being considered such as, interactive training of interviewers, special training modules, the possibility of conducting re-interviews and better tracing tools. These procedures should make it possible to make better use of the flexibility resulting from the automation of the process.

The case management system is also being redeveloped. One major consideration here is to obtain a robust communications system, in which changes can be sent out uniformly with a replication capability. While we still hope to develop a computer system that will be used for many years, the current reality seems to suggest that CAI is likely to continue to evolve rapidly. One challenge, then, since the technology is changing quickly (one need only think of the Internet), is to develop a new system that is flexible, so as to allow for adaptations without requiring a complete overhaul.

ACKNOWLEDGEMENTS

The authors would like to thank the many people of Household Survey Methods, Social Survey Methods, Household Surveys and Survey Operations Divisions who have contributed to the development of CAI at Statistics Canada over the years. It is their work that has made this paper possible. They would also like to thank Ann Brown, Brian Williams, Jean-Louis Tambay and Frank Mayda for their valuable comments that helped improve the quality of the paper.

REFERENCES

ALLARD, B., BRISEBOIS, F., DUFOUR, J., and SIMARD, M. (1996). How Do Interviewers Do Their job? A Look at New Data Quality Measures for the Canadian Labour Force Survey. Presented at the International Conference on Computer-assisted Survey Information Collection.

ALLARD, B., DUFOUR, J., SIMARD, M., and BASTIEN, J.-F. (1996). Pourquoi refuse-t-on de participer aux enquêtes? Le cas de l'Enquête sur la population active. Methodology Branch Working Paper, HSMD, 96-003F. Statistics Canada.

BRISEBOIS, F., DUFOUR, J., and LÉVESQUE, I. (1997). New LFS quality measures. Methodology Branch Working Paper, to be published. Statistics Canada.

BRODEUR, M., MONTIGNY, G., and BÉRARD, H. (1995). Challenge in developing the National Longitudinal Survey of Children. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-28.

BROWN, A., HALE, A., and MICHAUD, S. (1997). Use of Computer-assisted Interviewing in Longitudinal Surveys. Presented at the International Conference on Computer-assisted Survey Information Collection.

CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and data quality. In *Telephone Survey Methodology*, edited by R.M. Groves *et al.*, New York: John Wiley and Sons.

CATLIN, G., ROBERTS, K. and INGRAM S. (1996). The validity of self-reported chronic conditions in the National Population Health Survey. Presented at Symposium 96, Nonsampling Errors, Statistics Canada.

CLARK, C., MARTIN, J., and BATES, N. (1997). Development and Implementation of CASIC in Government Statistical Agencies. Presented at the International Conference on Computer-assisted Survey Information Collection.

DIBBS, R., HALE, A., LOVEROCK, R., and MICHAUD, S. (1995). Some Effects of Computer-assisted Interviewing on the Data Quality of the Survey of Labour and Income Dynamics. Survey of Labour and Income Dynamics Research Paper, 95-07. Statistics Canada.

DREW, D., GAMBINO, J., AKYEAMPONG, E., and WILLIAMS, B. (1991). Plans for the 1991 redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DUFOUR, J., KAUSHAL, R., CLARK, C., and BENCH, J. (1995). Converting the Labour Force Survey to Computer-assisted Interviewing. Methodology Branch Working Paper, HSMD, 95-009E. Statistics Canada.

DUFOUR, J., SIMARD, M., and MAYDA, F. (1995). The First Year of Computer-assisted Interviewing for the Canadian Labour Force Survey: An Update. Methodology Branch Working Paper, HSMD, 95-011E. Statistics Canada.

HALE, A., and MICHAUD, S. (1995). Dependent Interviewing: Impact on Recall and on Labour Market Transitions. Survey of Labour and Income Dynamics Research Paper, 95-06. Statistics Canada.

HIEMSTRA, D., LAVIGNE, M., and WEBBER, M. (1993). Labour Force Classification in SLID: Evaluation of Test 3A Results. Survey of Labour and Income Dynamics Research Paper, 93-14. Statistics Canada.

KAUSHAL, R., and LANIEL, N. (1995). Computer-assisted interviewing data quality test. *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 513-524.

LAVIGNE, M., and MICHAUD, S. (1995). Aspects généraux de l'Enquête sur la dynamique du travail et du revenu. *Recueil des textes des présentations du colloque sur les applications de la statistique*. L'association canadienne française pour l'avancement des sciences.

LYBERG, L., BIEMER, P., COLLINS, M., de LEEUW, E., DIPPO, C., SCHWARZ, N., and TREWIN, D. (1997). *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

MICHAUD, S., LE PETIT, C., and LAVIGNE, M. (1993). Qualitative Aspects of SLID Test 3A Data Collection. Survey of Labour and Income Dynamics Research Papers, 93-07. Statistics Canada

MICHAUD, S., LAVIGNE, M., and POTTLE, J. (1993). Qualitative Aspects of SLID Test 3B Data Collection. Survey of Labour and Income Dynamics Research Papers, 93-11. Statistics Canada.

MURRAY T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1990). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1990 Annual Research Conference.* U.S. Bureau of the Census.

NICHOLLS II, W.L., and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I. *Journal of Official Statistics*, 2, 93-115.

SIMARD, M., and DUFOUR, J. (1995). Impact of The Introduction of Computer-assisted Interviewing as the New Labour Force Survey Data Collection Method. Technical Report, Household Survey Methods Division, Statistics Canada.

SIMARD, M., DUFOUR, J., and MAYDA, F. (1995). The first year of computer-assisted interviewing as the Canadian Labour Force Survey data collection method. *Proceedings of Section on Survey Research Methods, American Statistical Association*, 533-538.

SINGH, M.P., GAMBINO, J., and LANIEL, N. (1993). Research studies for the Labour Force Survey sample redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

STATISTICS CANADA (1998). *Methodology of the Canadian Labour Force Survey*. Catalogue 71-526. To appear.

TAMBAY, J.-L., and CATLIN, G. (1995). Sample Design of the National Population Health Survey. Health Reports. Catalogue: 82-003, Statistics Canada. 7, 29-38.

WILLIAMS, B., and SPAULL, M. (1992). Computer-assisted Personal Interviewing LFS Datellite Test 0691-1191. Internal report. ISS Managers Conference, Statistics Canada.

Survey Methodology, December 1997
Vol. 23, No. 2, pp. 157-165
Statistics Canada

Regression Analysis of Data Files that are Computer Matched - Part II

FRITZ SCHEUREN and WILLIAM E. WINKLER[1]

ABSTRACT

Many policy decisions are best made when there is supporting statistical evidence based on analyses of appropriate microdata. Sometimes all the needed data exist but reside in multiple files for which common identifiers (*e.g.*, SIN's, EIN's, or SSN's) are unavailable. This paper demonstrates a methodology for analyzing two such files: (1) when there is common nonunique information subject to significant error and (2) when each source file contains noncommon quantitative data that can be connected with appropriate models. Such a situation might arise with files of businesses only having difficult-to-use name and address information in common, one file with the energy products consumed by the companies, and the other file containing the types and amounts of goods they produce. Another situation might arise with files on individuals in which one file has earnings data, another information about health-related expenses, and a third information about receipts of supplemental payments. The goal of the methodology presented is to produce valid statistical analyses; appropriate microdata files may or may not be produced.

KEY WORDS: Edit; Imputation; Record linkage; Regression analysis.

1. INTRODUCTION

1.1 Application Setting

To model the energy economy properly, an economist might need company-specific microdata on the fuel and feedstocks used by companies that are only available from Agency A and corresponding microdata on the goods produced for companies that is only available from Agency B. To model the health of individuals in society, a demographer or health science policy worker might need individual-specific information on those receiving social benefits from Agencies B1, B2, and B3, corresponding income information from Agency I, and information on health services from Agencies H1 and H2. Such modeling is possible if analysts have access to the microdata and if unique, common identifiers are available (*e.g.*, Oh and Scheuren 1975; Jabine and Scheuren 1986). If the only common identifiers are error-prone or nonunique or both, then probabilistic matching techniques (*e.g.*, Newcombe, Kennedy, Axford and James 1959, Fellegi and Sunter 1969) are needed.

1.2 Relation to Earlier Work

In earlier work (Scheuren and Winkler 1993), we provided theory showing that elementary regression analyses could be accurately adjusted for matching error, employing knowledge of the quality of the matching. In that work we relied heavily on an error-rate estimation procedure of Belin and Rubin (1995). In later research *e.g.*, (Winkler and Scheuren 1995, 1996), we showed that we could make further improvements by using noncommon quantitative data from the two files to improve matching and adjust statistical analyses for matching error. The main requirement – even in heretofore seemingly impossible situations – was that there exist a reasonable model for the relationships among the noncommon quantitative data. In the empirical example of this paper, we use data for which a very small subset of pairs can be accurately matched using name and address information only and for which the noncommon quantitative data is at least moderately correlated. In other situations, researchers might have a small microdata set that accurately represents relationships of noncommon data across a set of large administrative files or they might just have a reasonable guess at what the relationships among the noncommon data are. We are not sure, but conjecture that, with a reasonable starting point, the methods discussed here will succeed often enough to be of general value.

1.3 Basic Approach

The intuitive underpinnings of our methods are based on now well-known probabilistic record linkage (RL) and edit/imputation (EI) technologies. The ideas of modern RL were introduced by Newcombe (Newcombe *et al.* 1959) and mathematically formalized by Fellegi and Sunter (1969). Recent methods are described in Winkler (1994, 1995). EI has traditionally been used to clean up erroneous data in files. The most pertinent methods are based on the EI model of Fellegi and Holt (1976).

To adjust a statistical analysis for matching error, we employ a four-step recursive approach that is very powerful. We begin with an enhanced RL approach (*e.g.*, Winkler 1994, Belin and Rubin 1995) to delineate a subset of pairs of records in which the matching error rate is estimated to be very low. We perform a regression analysis, RA, on the

[1] Fritz Scheuren, Ernst and Young, 1225 Connecticut Avenue, N.W., Washington, DC 20036, U.S.A., Scheuren@aol.com; William E. Winkler, U.S. Bureau of the Census, Washington, DC 20023, U.S.A.

low-error-rate linked records and partially adjust the regression model on the remainder of the pairs by applying previous methods (Scheuren and Winkler 1993). Then, we refine the EI model using traditional outlier-detection methods to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (RA) is done and this time the results are fed back into the linkage step so that the RL step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, these *analytic linking* methods take the form

$$\nearrow \mathrm{RA} \searrow$$
$$\mathrm{RL} \leftarrow \mathrm{RA} \leftarrow \mathrm{EI}$$

1.4 Structure of What Follows

Beginning with this introduction, the paper is divided into five sections. In the second section, we undertake a short review of Edit/Imputation (EI) and Record Linkage (RL) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (RA) is so well known, our treatment of it is covered only in the particular simulated application (Section 3). The intent of these simulations is to use matching scenarios that are more difficult than what most linkers typically encounter. Simultaneously, we employ quantitative data that is both easy to understand but hard to use in matching. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

2. EI AND RL METHODS REVIEWED

2.1 Edit/Imputation

Methods of editing microdata have traditionally dealt with logical inconsistencies in data bases. Software consisted of if-then-else rules that were data-base-specific and very difficult to maintain or modify, so as to keep current. Imputation methods were part of the set of if-then-else rules and could yield revised records that still failed edits. In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976) introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values, so that the revised record satisfies all edits. An additional advantage of Fellegi and Holt (1976) systems is that their edit methods tie directly with current methods of imputing microdata (*e.g.*, Little and Rubin 1987).

Although we will only consider continuous data in this paper, EI techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$c_1 X < Y < c_2 X$$

In words,

> Y can be expected to be greater than $c_1 X$ and less than $c_2 X$; hence, if Y less than $c_1 X$ and greater than $c_2 X$, then the data record should be reviewed (with resource and other practical considerations determining the actual bounds used).

Here Y may be total wages, X the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$. When an (X, Y) pair associated with a record fails an edit, we may replace, say, Y with an estimate (or prediction).

2.2 Record Linkage

A record linkage process attempts to classify pairs in a product space $A \times B$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.* 1959; Newcombe, Fair and Lalonde 1992), Fellegi and Sunter (1969) considered ratios R of probabilities of the form

$$R = \Pr((\gamma \in \Gamma \mid M) / \Pr((\gamma \in \Gamma \mid U)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called *matching variables*. The decision rule is given by

> If $R > Upper$, then designate pair as a link.
>
> If $Lower \le R \le Upper$, then designate pair as a possible link and hold for clerical review.
>
> If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (typically a logarithm) a *matching weight* or *total agreement weight.*

With the availability of inexpensive computing power, there has been an outpouring of new work on record linkage techniques (*e.g.*, Jaro 1989, Newcombe, *et al.* 1992, Winkler 1994, 1995). The new computer-intensive methods reduce, or even sometimes eliminate, the need for clerical review when name, address, and other information used in matching is of reasonable quality. The proceedings from a recently concluded international conference on record linkage showcase these ideas and might be the best single reference (Alvey and Jamerson 1997).

3. SIMULATION SETTING

3.1 Matching Scenarios

For our simulations, we considered a scenario in which matches are virtually indistinguishable from nonmatches. In our earlier work (Scheuren and Winkler 1993), we considered three matching scenarios in which matches are more easily distinguished from nonmatches than in the scenario of the present paper.

In both papers, the basic idea is to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. Because the methods of this paper work better than what we did earlier, we only consider a matching scenario that we label "Second Poor," because it is more difficult than the poor (most difficult) scenario we considered previously.

We started here with two population files (sizes 12,000 and 15,000), each having good matching information and for which true match status was known. Three settings were examined: high, medium and low – depending on the extent to which the smaller file had cases also included in the larger file. In the high file inclusion situation, about 10,000 cases are on both files for a file inclusion or intersection rate on the smaller or base file of about 83%. In the medium file intersection situation, we took a sample of one file so that the intersection of the two files being matched was approximately 25%. In the low file intersection situation, we took samples of both files so that the intersection of the files being matched was approximately 5%. The number of intersecting cases, obviously, bounds the number of true matches that can be found.

We then generated quantitative data with known distributional properties and adjoined the data to the files. These variations are described below and displayed in Figure 1 where we show the poor scenario (labeled "first poor") of our previous 1993 paper and the "second poor" scenario used in this paper. In the figure, the match weight, the logarithm of R, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o).

3.2 "First Poor Scenario" (Figure 1a)

The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names in one of the files. Moderately severe typographical errors were made independently in one fourth of the addresses of the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was for the links to be made in a manner that a practitioner might choose after gaining only a little experience. The situation is analogous to that of using administrative lists of individuals where information used in matching is of poor quality. The true mismatch rate here was 10.1%.

3.3 "Second Poor" Scenario (Figure 1b)

The second poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names in one of the files. Severe typographical errors were made in one fourth of the addresses in the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. Name information – a key identifying characteristic – is often very difficult to compare effectively with business lists. The true mismatch rate was 14.6%.

3.4 Summary of Matching Scenarios

Clearly, depending on the scenario, our ability to distinguish between true links and true nonlinks differs significantly. With the first poor scenario, the overlap, shown visually between the log-frequency-versus-weight curves, is substantial (Figure 1a); and, with the second poor scheme, the overlap of the log-frequency-versus-weight curves is almost total (Figure 1b). In the earlier work, we showed that our theoretical adjustment procedure worked well using the known true match rates in our data sets. For situations where the curves of true links and true nonlinks were reasonably well separated, we accurately estimated error rates via a procedure of Belin and Rubin (1995) and our procedure could be used in practice. In the poor matching scenario of that paper (first poor scenario of this paper), the Belin-Rubin procedure was unable to provide accurate estimates of error rates but our theoretical adjustment procedure still worked well. This indicated that we either had to find an enhancement to the Belin-Rubin procedures or to develop methods that used more of the available data. (That conclusion, incidentally, from our earlier workled, after some false starts, to the present approach.)

3.5 Quantitative Scenarios

Having specified the above linkage situations, we used SAS to generate ordinary least squares data under the model $Y = 6X + \varepsilon$. The X values were chosen to be uniformly distributed between 1 and 101. The error terms, are normal and homoscedastic with variances 13,000, 36,000, and 125,000, respectively. The resulting regressions of Y on X have R^2 values in the true matched population of 70%, 47%, and 20%, respectively. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. To make modeling and analysis even more difficult in the high file overlap scenario, we used all false matches and only 5% of the true matches; in the medium file overlap scenario, we used all false matches and only 25% of true matches. (Note: Here to heighten the visual effect, we have introduced another random sampling step, so the reader can "see"

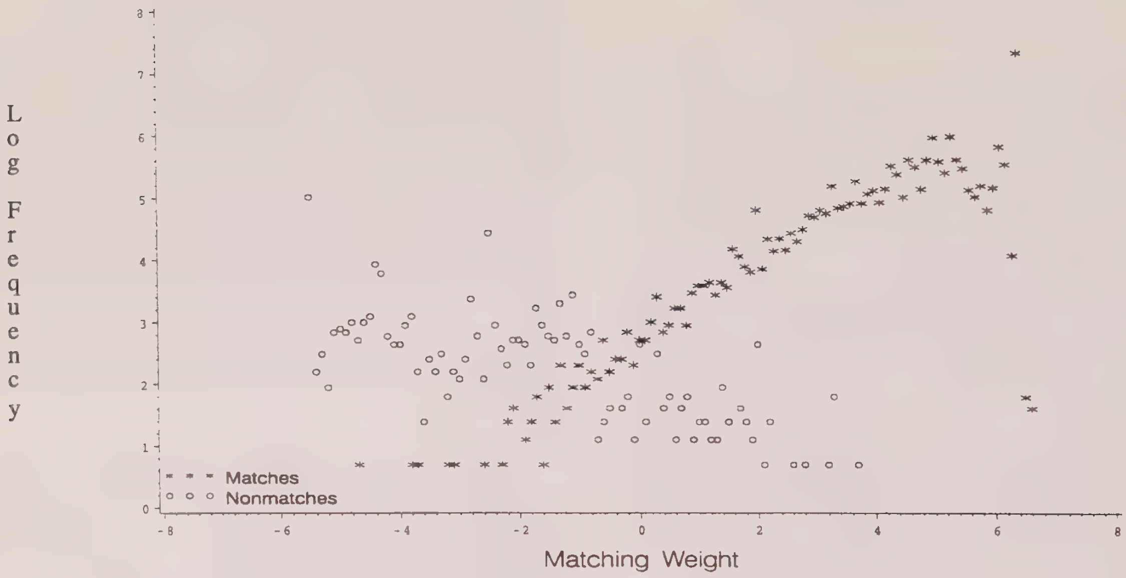


Figure 1a. 1st Poor Matching Scenario

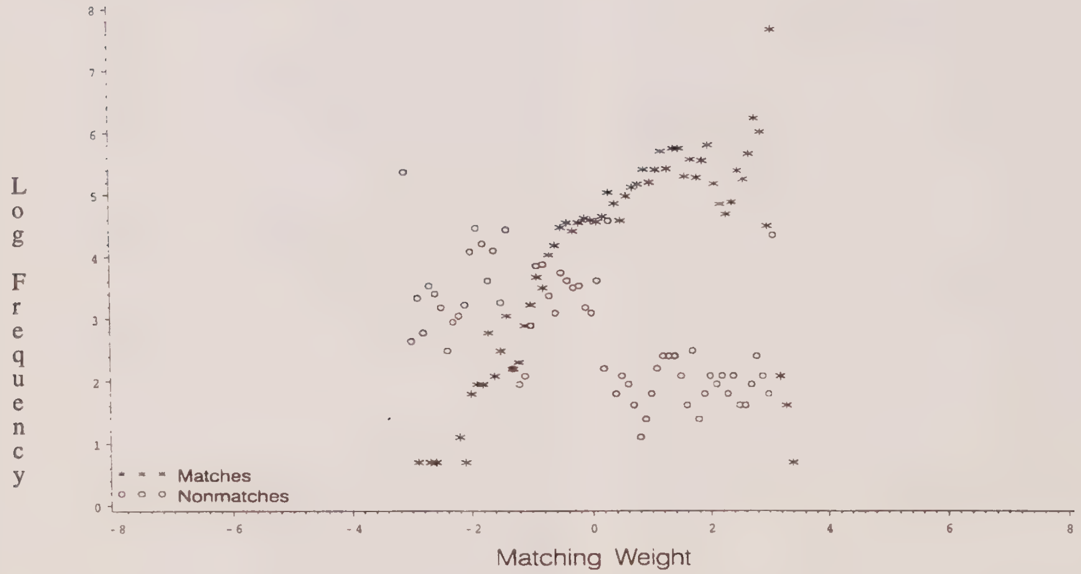


Figure 1b. 2nd Poor Matching Scenario

better in the figures the effect of bad matching. This sample depends on the match status of the case and is confined only to those cases that were matched, whether correctly or falsely.)

A crucial practical assumption for the work of this paper is that analysts are able to produce a reasonable model (guesstimate) for the relationships between the noncommon quantitative items. For the initial modeling in the empirical example of this paper, we use the subset of pairs for which matching weight is high and the error-rate is low. Thus, the number of false matches in the subset is kept to a minimum. Although neither the procedure of Belin and Rubin (1995) nor an alternative procedure of Winkler (1994), that requires an *ad hoc* intervention, could be used to estimate error rates, we believe it is possible for an experienced matcher to pick out a low-error-rate set of pairs even in the second poor scenario.

4. SIMULATION RESULTS

Most of this Section is devoted to presenting graphs and results of the overall process for the second poor scenario, where the R^2 value is moderate, and the intersection between the two files is high. These results best illustrate the procedures of this paper. At the end of the Section (in subsection 4.8), we summarize results over all R^2 situations and all overlaps. To make the modeling more difficult and show the power of the analytic linking methods, we use all false matches and a random sample of only 5% of the true matches. We only consider pairs having matching weight above a lower bound that we determine based on analytic considerations and experience. For the pairs of our analysis, the restriction causes the number of false matches to significantly exceed the number of true matches. (Again, this is done to heighten the visual effect of matching failures and to make the problem even more difficult.)

To illustrate the data situation and the modeling approach, we provide triples of plots. The first plot in the triple shows the true data situation as if each record in one file was linked with its true corresponding record in the other file. The quantitative data pairs correspond to the truth. In the second plot, we show the observed data. Where many of the pairs are in error because they correspond to false matches. To get to the third plot in the triple, we model using a small number of pairs (approximately 100) and then replace outliers with pairs in which the observed Y-value is replaced with a predicted Y-value.

4.1 Initial True Regression Relationship

In Figure 2a, the actual true regression relationship and related scatterplot are shown, for one of our simulations, as they would appear if there were no matching errors. In this figure and the remaining ones, the true regression line is always given for reference. Finally, the true population slope or *beta* coefficient (at 5.85) and the R^2 value (at 43%) are provided for the data (sample of pairs) being displayed.

4.2 Regression After Initial RL→RA Step

In Figure 2b, we are looking at the regression on the actual observed links – not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between Y and X. The observed slope or *beta* coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected – falling to 7% from 43%.

4.3 Regression After First Combined RL→RA→EI→RA Step

Figure 2c completes our display of the first cycle of the iterative process we are employing. Here we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00 or larger, an attempt was made to improve the poor results given in Figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 460 or more were removed and the regression refit on the remaining pairs. This new equation, used in Figure 2c, was essentially $Y = 4.78X + \varepsilon$, with a variance of 40,000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the estimated *beta* coefficient from 4.78 to 5.4. If a pair of matched records yielded an outlier, then predicted values (not shown) using the equation $Y = 5.4X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.4 Second True Reference Regression

Figure 3a displays a scatterplot of X and Y as they would appear if they could be true matches based on a second RL step. Note here that we have a somewhat different set of linked pairs this time from earlier, because we have used the regression results to help in the linkage. In particular, the second RL step employed the predicted Y values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second RL step. Since a considerably better link was obtained, there were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1,104 in Figures 2a through 2c to 650 for Figures 3a through 3c. In this second iteration, the true slope or *beta* coefficient and the R^2 values remained, though, virtually identical for the estimated slope (5.85 v. 5.91) and fit (43% v. 48%).

4.5 Regression After Second RL→RA Step

In Figure 3b, we see a considerable improvement in the relationship between Y and X using the actual observed links after the second RL step. The estimated slope has risen from 2.47 initially to 4.75 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 33%.

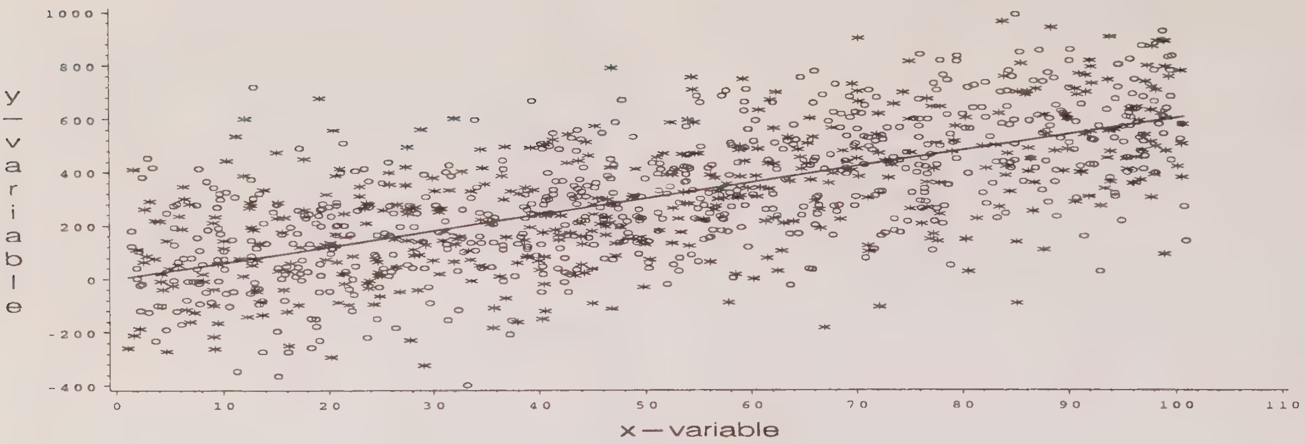


Figure 2a. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, True Data, HighOverlap,
1104 Points, beta = 5.85, R–square = 0.43

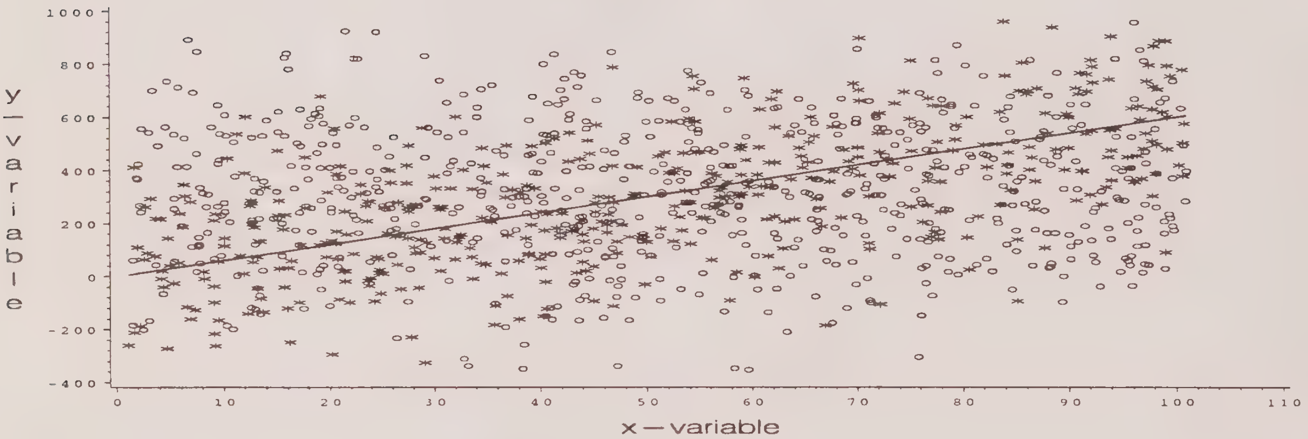


Figure 2b. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Observed Data, HighOverlap,
1104 Points, beta = 2.47, R – square = 0.07

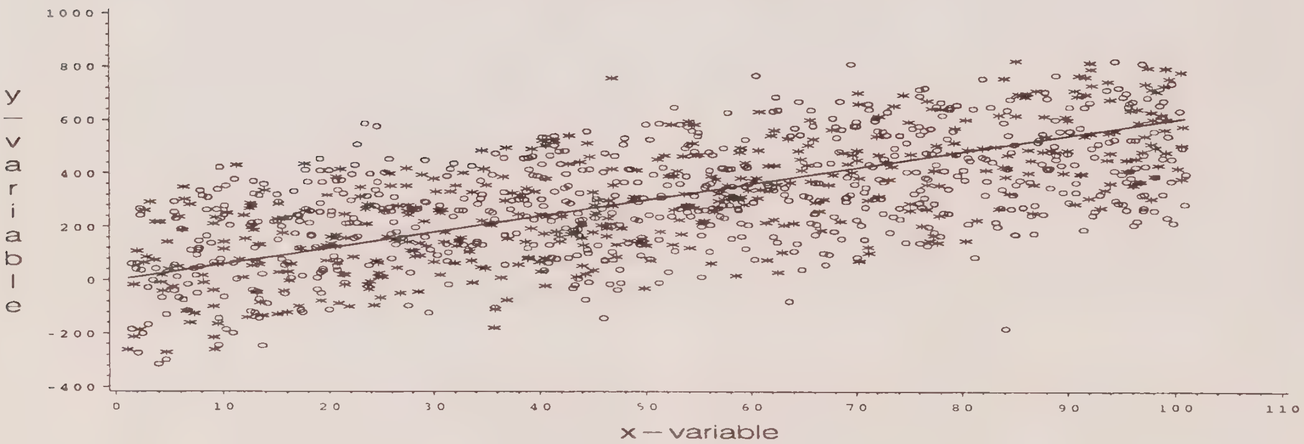


Figure 2c. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Outlier – Adjusted Data
1104 Points, beta = 4.78, R – square = 0.40

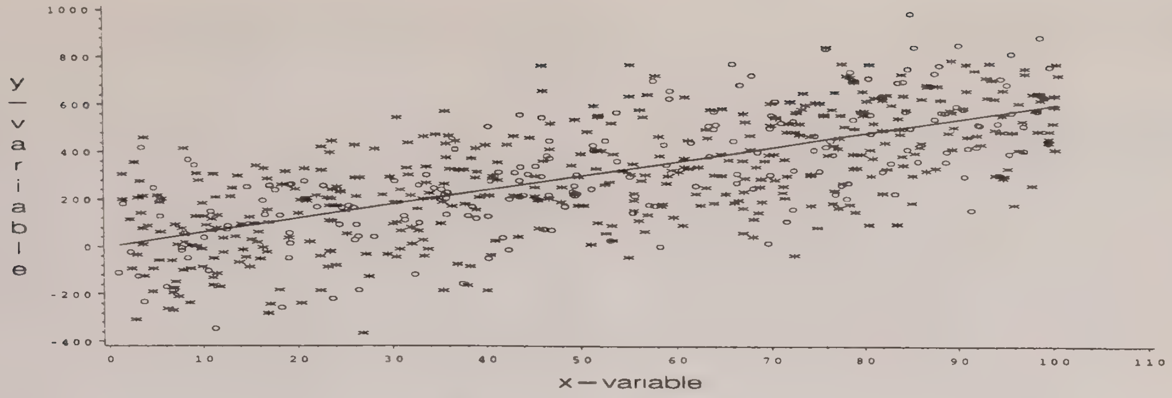


Figure 3a. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, True Data, HighOverlap,
650 Points, beta = 5.91 R – square = 0.48

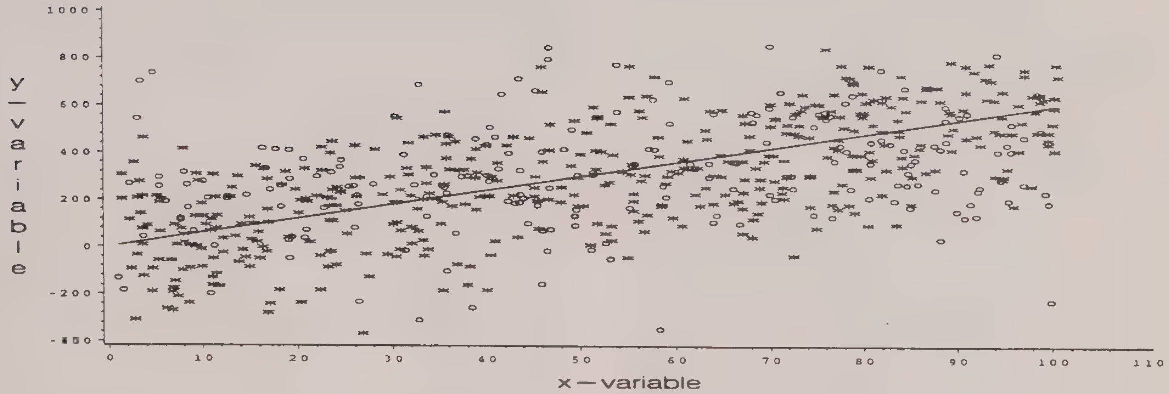


Figure 3b. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Observed Data, HighOverlap
650 Points, beta = 4.75, R – square = 0.33

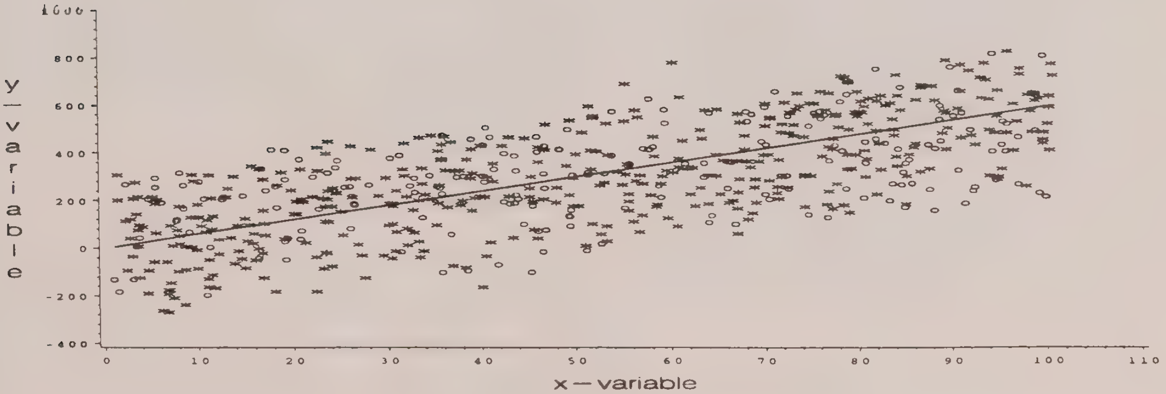


Figure 3c. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Outlier – Adjusted Data
650 Points, beta = 5.26, R – square = 0.47

4.6 Regression After Second Combined RL→RA→EI→RA Step

Figure 3c completes the display of the second cycle of our iterative process. Here we have edited the data as follows. Using the fit (from subsection 4.5), another set of predicted values was obtained for all the matched cases (as in subsection 4.3). This new equation was essentially $Y = 5.26X + \varepsilon$, with a variance of about 35,000. If a pair of matched records yields an outlier, then predicted values using the equation $Y = 5.3X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.7 Additional Iterations

While we did not show it in this paper, we did iterate through a third matching pass. The *beta* coefficient, after adjustment, did not change much. We do not conclude from this that asymptotic unbiasedness exists; rather that the method, as it has evolved so far, has a positive benefit and that this benefit may be quickly reached.

4.8 Further Results

Our further results are of two kinds. We looked first at what happened in the medium R^2 scenario (*i.e.*, R^2 equal to .47) for the medium- and low- file intersection situations. We further looked at the cases when R^2 was higher (at .70) or lower (at .20). For the medium R^2 scenario and low intersection case the matching was somewhat easier. This occurs because there were significantly fewer false-match candidates and we could more easily separate true matches from false matches. For the high R^2 scenarios, the modeling and matching were also more straightforward than they were for the medium R^2 scenario. Hence, there were no new issues there either.

On the other hand, for the low R^2 scenario, no matter what degree of file intersection existed, we were unable to distinguish true matches from false matches, even with the improved methods we are using. The reason for this, we believe, is that there are many outliers associated with the true matches. We can no longer assume, therefore, that a moderately higher percentage of the outliers in the regression model are due to false matches. In fact, with each true match that is associated with an outlier Y-value, there may be many false matches that have Y-values that are closer to the predicted Y-value than the true match.

5. COMMENTS AND FUTURE STUDY

5.1 Overall Summary

In this paper, we have looked at a very restricted analysis setting: a simple regression of one quantitative dependent variable from one file matched to a single quantitative independent variable from another file. This standard analysis was, however, approached in a very nonstandard setting. The matching scenarios, in fact, were quite challenging. Indeed, just a few years ago, we might have said that the "second poor" matching scenario appeared hopeless.

On the other hand, as discussed below, there are many loose ends. Hence, the demonstration given here can be considered, quite rightly in our view, as a limited accomplishment. But make no mistake about it, we are doing something entirely new. In past record linkage applications, there was a clear separation between the identifying data and the analysis data. Here, we have used a regression analysis to improve the linkage and the improved linkage to improve the analysis and so on.

Earlier, in our 1993 paper, we advocated that there be a unified approach between the linkage and the analysis. At that point, though, we were only ready to propose that the linkage probabilities be used in the analysis to correct for the failures to complete the matching step satisfactorily. This paper is the first to propose a completely unified methodology and to demonstrate how it might be carried out.

5.2 Planned Application

We expect that the first applications of our new methods will be with large business data bases. In such situations, noncommon quantitative data are often moderately or highly correlated and the quantitative variables (both predicted and observed) can have great distinguishing power for linkage, especially when combined with name information and geographic information, such as a postal (*e.g.*, ZIP) code.

A second observation is also worth making about our results. The work done here points strongly to the need to improve some of the now routine practices for protecting public use files from reidentification. In fact, it turns out that in some settings – even after quantitative data have been confidentiality protected (by conventional methods) and without any directly identifying variables present – the methods in this paper can be successful in reidentifying a substantial fraction of records thought to be reasonably secure from this risk (as predicted in Scheuren 1995). For examples, see Winkler (1997).

5.3 Expected Extensions

What happens when our results are generalized to the multiple regression case? We are working on this now and results are starting to emerge which have given us insight into where further research is required. We speculate that the degree of underlying association R^2 will continue to be the dominant element in whether a usable analysis is possible.

There is also the case of multivariate regression. This problem is harder and will be more of a challenge. Simple multivariate extensions of the univariate comparison of Y values in this paper have not worked as well as we would like. For this setting, perhaps, variants and extensions of Little and Rubin (1987, Chapters 6 and 8) will prove to be a good starting point

5.4 "Limited Accomplishment"

Until now an analysis based on the second poor scenario would not have been even remotely sensible. For this reason alone we should be happy with our results. A closer examination, though, shows a number of places where the approach demonstrated is weaker than it needs to be or simply unfinished. For those who want theorems proven, this may be a particularly strong sentiment. For example, a convergence proof is among the important loose ends to be dealt with, even in the simple regression setting. A practical demonstration of our approach with more than two matched files also is necessary, albeit this appears to be more straightforward.

5.5 Guiding Practice

We have no ready advise for those who may attempt what we have done. Our own experience, at this point, is insufficient for us to offer ideas on how to guide practice, except the usual extra caution that goes with any new application. Maybe, after our own efforts and those of others have matured, we can offer more.

REFERENCES

ALVEY, W., and JAMERSON, B. (Eds.) (1997). *Record Linkage Techniques – 1997.* Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, Arlington, VA.

BELIN, T.R., and RUBIN, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

FELLEGI, I., and HOLT, T. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

FELLEGI, I., and SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

JABINE, T.B., and SCHEUREN, F. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 89, 414-420.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis With Missing Data.* New York: John Wiley.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science,* 130, 954-959.

NEWCOMBE, H., FAIR, M., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.

OH, H.L., and SCHEUREN, F. (1975). Fiddling around with mismatches and nonmatches. *Proceedings of the Social Statistics Section, American Statistical Association.*

SCHEUREN, F. (1995). Review of private lives and public policies: Confidentiality and accessibility of government services. *Journal of the American Statistical Association*, 90, 386-387.

SCHEUREN, F., and WINKLER, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

WINKLER, W.E. (1994). Advanced methods of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.

WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox *et al.*). New York: John Wiley, 355-384.

WINKLER, W.E., and SCHEUREN, F. (1995). Linking data to create information. *Proceedings: Symposium 95, From Data to Information-Methods and Systems,* Statistics Canada, 29-37.

WINKLER, W.E., and SCHEUREN, F. (1996). Recursive analysis of linked data files. *Proceedings of the 1996 Annual Research Conference.* U.S. Bureau of the Census.

WINKLER, W.E. (1997). Producing Public-Use Microdata That are Analytically Valid and Confidential. Presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1997. An asterisk indicates that the person served more than once.

J.C. Arnold, *Virginia Polytechnic Institute*
M. Bankier, *Statistics Canada*
* D.R. Bellhouse, *University of Western Ontario*
* T.R. Belin, *University of California - Los Angeles*
* D.A. Binder, *Statistics Canada*
G.J. Brackstone, *Statistics Canada*
F.J. Breidt, *Iowa State University*
A. Brinkley, *U.S. Bureau of the Census*
L. Cahoon, *U.S. Bureau of the Census*
N. Caron, *Institut national de la statistique et des études économiques*
R. Caspar, *Research Triangle Institute*
R. Chambers, *University of Southampton*
S.X. Chen, *New York University*
G.H. Choudhry, *Statistics Canada*
W. Davis, *Klemm Analysis Group*
* J. Denis, *Statistics Canada*
J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistics Canada*
J.D. Drew, *Statistics Canada*
D.F. Findlay, *U.S. Bureau of the Census*
B. Forsyth, *Westat, Inc.*
L.A. Franklin, *Indiana State University*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
G. Gates, *U.S. Bureau of the Census*
B.V. Greenberg, *U.S. Bureau of the Census*
* R.M. Groves, *University of Maryland*
J.-P. Gwet, *Westat, Inc.*
* M.A. Hidiroglou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
C. Julien, *Statistics Canada*
* G. Kalton, *Westat, Inc.*
S. Kaufman, *National Center for Education Statistics*
D. Kerr, *Statistics Canada*
J.J. Kim, *U.S. Bureau of the Census*
P. Kokic, *University of Southampton*
M. Kovacevic, *Statistics Canada*
R. Lachapelle, *Statistics Canada*
M. Latouche, *Statistics Canada*
* P. Lavallée, *Statistics Canada*
J. Ledent, *Université de Québec*
S. Linacre, *Australian Bureau of Statistics*
R. Little, *University of Michigan*
D. Malec, *National Center for Health Statistics*
* H. Mantel, *Statistics Canada*
N. Mathiowetz, *University of Maryland*
C. Moriarity, *National Center for Health Statistics*
* B. Nandram, *Worcester Polytechnic Institute*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
* B. Quenneville, *Statistics Canada*
T.E. Raghunathan, *University of Michigan*
E. Rancourt, *Statistics Canada*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
* I. Sande, *Bell Communications Research, U.S.A.*
G. Sande, *Sande & Assoc.*
F.J. Scheuren, *George Washington University*
* J. Sedransk, *Case Western Reserve University*
J. Shao, *University of Wisconsin - Madison*
* A.C. Singh, *Statistics Canada*
* M.P. Singh, *Statistics Canada*
B.K. Sinha, *University of Maryland*
* R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
G. Smith, *Statistics Canada*
P. Steel, *U.S. Bureau of the Census*
* D. Stukel, *Statistics Canada*
W. Sun, *Statistics Canada*
J.-L Tambay, *Statistics Canada*
A. Théberge, *Statistics Canada*
* R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
I. Thomsen, *Statistics Norway*
Y. Tillé, *École nationale de statistique et de l'analyse de l'information*
R. Valliant, *U.S. Bureau of Labor Statistics*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
F. Yu, *Australian Bureau of Statistics*
M. Yu, *Statistics Canada*
* A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1997 issues: S. Beauchamp and L. Durocher (Composition Unit) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge D. Blair, S. DiLoreto, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 4, December/décembre 1997

Christian GENEST
Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995

Debajyoti SINHA
Time-discrete beta process model for interval-censored survival data

Lynn KUO and Bani MALLICK
Bayesian semiparamametric inference for the accelerated failure time model

Stephen G. WALKER and Bani K. MALLICK
A note on the scale parameter of the Dirichlet process

Nancy HECKMAN and John RICE
Line transects of two dimensional random fields: Estimation and design

Fulvio DE SANTIS and Fulvio SPEZZAFERRI
Alternative Bayes factors for model selection

Gemai CHEN and Richard A. LOCKHART
Box-Cox transformed linear models: A parameter based asymptotic approach

Holger DETTE
E-optimal designs for regression models with quantitative factors - a reasonable choice?

Jeesen CHEN
A general lower bound of minimax risk for absolute error loss

Yodit SEIFU and N. REID
Applications of bivariate and univariate local Lyapunov exponents

Robert TIBSHIRANI and Donald A. REDELMEIER
Cellular telephones and motor vehicle collisions: some variations on matched pairs analysis

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 13, Number 4, 1997

A Sampling Scheme With Partial Replacement
J.L. Sánchez-Crespo 327

Sources of Error in a Survey on Sexual Behavior
R. Tourangeau, K. Rasinski, J.B. Jobe, T.W. Smith, and W.F. Pratt 341

Developing an Estimation Strategy for a Pesticide Data Program
Phillip S. Kott and D. Andrew Carr 367

Estimating Interpolated Percentiles from Grouped Data with Large Samples
Edward L. Korn, Douglas Midthune, and Barry I. Graubard 385

Ratio Estimation of Hardcore Drug Use
Doug Wright, Joe Gfroerer, and Joan Epstein 401

Statistical Disclosure Control and Sampling Weights
A.G. de Waal and L.C.R.J. Willenborg 417

Book Reviews 435

Editorial Collaborators 447

Index to Volume 13, 1997 449

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, nº 1) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 13, Number 4, 1997

A Sampling Scheme With Partial Replacement
J.L. Sánchez-Crespo 327

Sources of Error in a Survey on Sexual Behavior
R. Tourangeau, K. Rasinski, J.B. Jobe, T.W. Smith, and W.F. Pratt 341

Developing an Estimation Strategy for a Pesticide Data Program
Phillip S. Kott and D. Andrew Carr 367

Estimating Interpolated Percentiles from Grouped Data with Large Samples
Edward L. Korn, Douglas Midthune, and Barry I. Graubard 385

Ratio Estimation of Hardcore Drug Use
Doug Wright, Joe Gfroerer, and Joan Epstein 401

Statistical Disclosure Control and Sampling Weights
A.G. de Waal and L.C.R.J. Willenborg 417

Book Reviews 435

Editorial Collaborators 447

Index to Volume 13, 1997 449

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS TABLE DES MATIÈRES

Volume 25, No. 4, December/décembre 1997

Christian GENEST
Statistics on statistics: measuring research productivity by journal publications between 1985 and 1995

Debajyoti SINHA
Time-discrete beta process model for interval-censored survival data

Lynn KUO and Bani MALLICK
Bayesian semiparamametric inference for the accelerated failure time model

Stephen G. WALKER and Bani K. MALLICK
A note on the scale parameter of the Dirichlet process

Nancy HECKMAN and John RICE
Line transects of two dimensional random fields: Estimation and design

Fulvio DE SANTIS and Fulvio SPEZZAFERRI
Alternative Bayes factors for model selection

Gemai CHEN and Richard A. LOCKHART
Box-Cox transformed linear models: A parameter based asymptotic approach

Holger DETTE
E-optimal designs for regression models with quantitative factors - a reasonable choice?

Jeesen CHEN
A general lower bound of minimax risk for absolute error loss

Yodit SEIFU and N. REID
Applications of bivariate and univariate local Lyapunov exponents

Robert TIBSHIRANI and Donald A. REDELMEIER
Cellular telephones and motor vehicle collisions: some variations on matched pairs analysis

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1997. Un astérisque indique que la personne a participé plus d'une fois.

J.C. Arnold, *Virginia Polytechnic Institute*
M. Bankier, *Statistique Canada*
* D.R. Bellhouse, *University of Western Ontario*
* T.R. Belin, *University of California - Los Angeles*
* D.A. Binder, *Statistique Canada*
G.J. Brackstone, *Statistique Canada*
F.J. Breidt, *Iowa State University*
A. Brinkley, *U.S. Bureau of the Census*
L. Cahoon, *U.S. Bureau of the Census*
N. Caron, *Institut national de la statistique et des études économiques*
R. Caspar, *Research Triangle Institute*
R. Chambers, *University of Southampton*
S.X. Chen, *New York University*
G.H. Choudhry, *Statistique Canada*
W. Davis, *Klemm Analysis Group*
* J. Denis, *Statistique Canada*
J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistique Canada*
J.D. Drew, *Statistique Canada*
D.F. Findlay, *U.S. Bureau of the Census*
B. Forsyth, *Westat, Inc.*
L.A. Franklin, *Indiana State University*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
G. Gates, *U.S. Bureau of the Census*
B.V. Greenberg, *U.S. Bureau of the Census*
* R.M. Groves, *University of Maryland*
J.-P. Gwet, *Westat, Inc.*
* M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
C. Julien, *Statistique Canada*
* G. Kalton, *Westat, Inc.*
S. Kaufman, *National Center for Education Statistics*
D. Kerr, *Statistique Canada*
J.J. Kim, *U.S. Bureau of the Census*
P. Kokic, *University of Southampton*
M. Kovacevic, *Statistique Canada*
R. Lachapelle, *Statistique Canada*
M. Latouche, *Statistique Canada*
* P. Lavallée, *Statistique Canada*
J. Ledent, *Université de Québec*
S. Linacre, *Australian Bureau of Statistics*
R. Little, *University of Michigan*
D. Malec, *National Center for Health Statistics*
* H. Mantel, *Statistique Canada*
N. Mathiowetz, *University of Maryland*
C. Moriarity, *National Center for Health Statistics*
* B. Nandram, *Worcester Polytechnic Institute*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Pfeffermann, *Hebrew University*
* B. Quenneville, *Statistique Canada*
T.E. Raghunathan, *University of Michigan*
E. Rancourt, *Statistique Canada*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistique Canada*
* I. Sande, *Bell Communications Research, U.S.A.*
G. Sande, *Sande & Assoc.*
F.J. Scheuren, *George Washington University*
* J. Sedransk, *Case Western Reserve University*
J. Shao, *University of Wisconsin - Madison*
* A.C. Singh, *Statistique Canada*
* M.P. Singh, *Statistique Canada*
B.K. Sinha, *University of Maryland*
* R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
G. Smith, *Statistique Canada*
P. Steel, *U.S. Bureau of the Census*
* D. Stukel, *Statistique Canada*
W. Sun, *Statistique Canada*
J.-L Tambay, *Statistique Canada*
A. Théberge, *Statistique Canada*
* R. Thomas, *Carleton University*
M. Thompson, *University of Waterloo*
I. Thomsen, *Statistics Norway*
Y. Tillé, *École nationale de statistique et de l'analyse de l'information*
R. Valliant, *U.S. Bureau of Labor Statistics*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
F. Yu, *Australian Bureau of Statistics*
M. Yu, *Statistique Canada*
* A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1997: S. Beauchamp et L. Durocher (Unité de composition) et L. Perreault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à D. Blair, S. DiLoreto, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

devons-nous être satisfaits de nos résultats. Un examen plus approfondi révèle toutefois un certain nombre de lacunes, qui indiquent que l'approche illustrée est plus faible qu'elle ne devrait l'être ou qu'elle n'est tout simplement pas finie. Pour ceux qui recherchent une méthode par démonstration de théorèmes, ceci peut poser un problème particulièrement grand. La preuve de convergence, par exemple, est un des points importants à régler, même pour les cas de régression simple. Il nous faut également faire une démonstration pratique de notre approche sur plus de deux fichiers appariés, encore que cela puisse sembler plus simple.

5.5 Guide de pratique

Nous n'avons pour l'instant aucun conseil à formuler pour quiconque voudrait faire l'essai de notre approche. Notre expérience, à ce stade-ci, est en effet insuffisante pour que nous puissions formuler des idées sur la façon d'orienter la pratique, si ce n'est que de rappeler les précautions additionnelles usuelles qui s'imposent avec toute nouvelle application. Peut-être serons-nous en mesure de formuler d'autres conseils, lorsque nos propres efforts et ceux d'autres analystes auront mûri.

BIBLIOGRAPHIE

ALVEY, W., et JAMERSON, B. (Éds.) (1997). *Record Linkage Techniques – 1997.* Recueil du An International Record Linkage Workshop and Exposition, le 20-21 mars 1997, Arlington, VA.

BELIN, T.R., et RUBIN, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

FELLEGI, I., et HOLT, T. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

FELLEGI, I., et SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

JABINE, T.B., et SCHEUREN, F. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 89, 414-420.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis With Missing Data*. New York: John Wiley.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science,* 130, 954-959.

NEWCOMBE, H., FAIR, M., et LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.

OH, H.L., et SCHEUREN, F. (1975). Fiddling around with mismatches and nonmatches. *Proceedings of the Social Statistics Section, American Statistical Association.*

SCHEUREN, F. (1995). Review of private lives and public policies: Confidentiality and accessibility of government services. *Journal of the American Statistical Association*, 90, 386-387.

SCHEUREN, F., et WINKLER, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

WINKLER, W.E. (1994). Advanced methods of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.

WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Éds. B.G. Cox et coll.). New York: J. Wiley, 355-384.

WINKLER, W.E., et SCHEUREN, F. (1995). Couplage des données pour créer l'information. Recueil: *Symposium 95, Des données à l'information – méthodes et systèmes,* Statistique Canada, 31-40.

WINKLER, W.E., et SCHEUREN, F. (1996). Recursive analysis of linked data files. *Proceedings of the 1996 Annual Research Conference.* U.S. Bureau of the Census.

WINKLER, W.E. (1997). Producing Public-Use Microdata That are Analytically Valid and Confidential. Présenté au 1997 Joint Statistical Meetings, Anaheim, CA.

4.7 Itérations additionnelles

Bien que les résultats ne soient pas présentés ici, nous avons effectué un troisième cycle d'appariement. Le coefficient bêta, après ajustement, a peu changé. Nous n'en concluons pas à l'absence de biais asymptotique, mais présumons plutôt que la méthode – sous sa forme actuelle – comporte des avantages dont on peut rapidement tirer profit.

4.8 Autres résultats

Nos autres résultats sont de deux types. Nous avons d'abord examiné ce qu'il était arrivé avec le scénario moyen pour R^2 (c.-à-d. R^2 égal à 0,47), pour les cas d'intersection faible et modérée. Nous avons à nouveau examiné les cas où la valeur de R^2 était plus élevée (0,70) ou plus faible (0,20). Dans le cas du scénario moyen pour R^2 avec faible intersection, l'appariement a été légèrement plus facile, du fait qu'il y a eu beaucoup moins de faux appariements et qu'il a été plus facile de séparer les vrais appariements des appariements faux. Pour les scénarios avec fortes valeurs de R^2, la modélisation et l'appariement ont eux aussi été plus simples qu'avec le scénario moyen. Il n'y a donc pas eu de nouveaux problèmes là non plus.

À l'inverse, avec le scénario à faible valeur de R^2, il nous a été impossible de distinguer les appariements vrais des faux, quel que fut le degré d'intersection, et ce même avec nos méthodes améliorées. À notre avis, ceci est dû au nombre élevé de valeurs aberrantes associées aux appariements vrais. Nous ne pouvons donc plus présumer qu'un pourcentage modérément élevé de valeurs aberrantes dans le modèle de régression soit dû à des appariements faux. En fait, pour chaque appariement vrai associé à une valeur aberrante de Y, il peut y avoir bon nombre d'appariements faux dont les valeurs de Y se rapprochent davantage de la valeur prévue que l'appariement vrai.

5. COMMENTAIRES ET AUTRES ÉTUDES

5.1 Résumé

Nous avons utilisé dans cet article un cadre d'analyse très restreint, à savoir une régression simple d'une variable dépendante quantitative d'un fichier en fonction d'une variable indépendante quantitative d'un autre fichier. Cette analyse courante a toutefois été traitée dans un cadre très inhabituel et les scénarios d'appariement ont été très complexes. De fait, il y a à peine quelques années, le deuxième scénario d'appariement pauvre aurait sans doute semblé «sans espoir».

Cependant, comme nous l'expliquons ci-après, de nombreux aspects restent encore à régler. Aussi la démonstration présentée ici peut-elle être qualifiée – à juste titre croyons-nous – de réalisation limitée. Cependant, qu'on ne s'y méprenne pas, notre approche est tout à fait nouvelle. Auparavant, il y avait une nette séparation entre les données d'identification et les données d'analyse pour le couplage d'enregistrements. Ici, nous utilisons une analyse de régression pour obtenir un meilleur couplage et ce couplage amélioré sert à améliorer l'analyse, et ainsi de suite.

Dans notre article de 1993, nous préconisions une approche unifiée entre le couplage et l'analyse. À cette époque, toutefois, nous ne pouvions que proposer que les probabilités de couplage soient utilisées dans l'analyse pour corriger en fonction des rejets et permettre le parachèvement adéquat de l'étape d'appariement. Le présent article est le premier à proposer une méthode complètement unifiée et à démontrer comment l'appliquer.

5.2 Application prévue

Nous croyons que les premières applications de nos nouvelles méthodes porteront sur de larges bases de données d'entreprises, où les données quantitatives non communes sont souvent modérément ou fortement corrélées et où les variables quantitatives (à la fois prévues et observées) peuvent avoir un grand pouvoir distinctif pour le couplage, en particulier lorsqu'elles sont combinées à des informations sur le nom et le lieu géographique, comme le code postal.

Il est également une deuxième observation qu'il convient de faire au sujet de nos résultats. Ainsi, les travaux effectués à ce jour font largement ressortir la nécessité d'améliorer certaines techniques actuellement utilisées de routine pour protéger les fichiers à grande diffusion contre une ré-identification. En fait, il s'avère que, dans certaines situations – même après protection de la confidentialité des données quantitatives (selon les méthodes traditionnelles) et en l'absence de toute variable d'identification directe – les méthodes définies dans le présent document peuvent réussir à identifier de nouveau une fraction substantielle des enregistrements que l'on croyait raisonnablement à l'abri de ce risque (tel que prédit par Scheuren 1995). Pour des exemples, voir Winkler 1997.

5.3 Extensions prévues

Qu'advient-il lorsqu'il y a généralisation de nos résultats, dans les cas de régression multiple? Nous étudions actuellement ce phénomène et nos premiers résultats indiquent certains domaines sur lesquels devraient porter les recherches futures. Nous croyons que le degré d'association sous-jacente R^2 continuera d'être l'élément dominant quant à savoir si une analyse utilisable est possible.

Il y a également le cas de la régression à variables multiples, qui pose un problème plus difficile et exigeant. Dans ce document, les extensions multidimensionnelles simples de la comparaison à une variable des valeurs de Y n'ont pas donné les résultats espérés. Pour une telle analyse, il est possible que les variantes et extensions de Little et Rubin (1987, chapitres 6 et 8) constitueront un bon point de départ.

5.4 Réalisation «limitée»

Jusqu'à aujourd'hui, il aurait été absolument insensé de penser à faire une analyse basée sur le deuxième scénario pauvre. Aussi, même si ce n'est que pour cette raison,

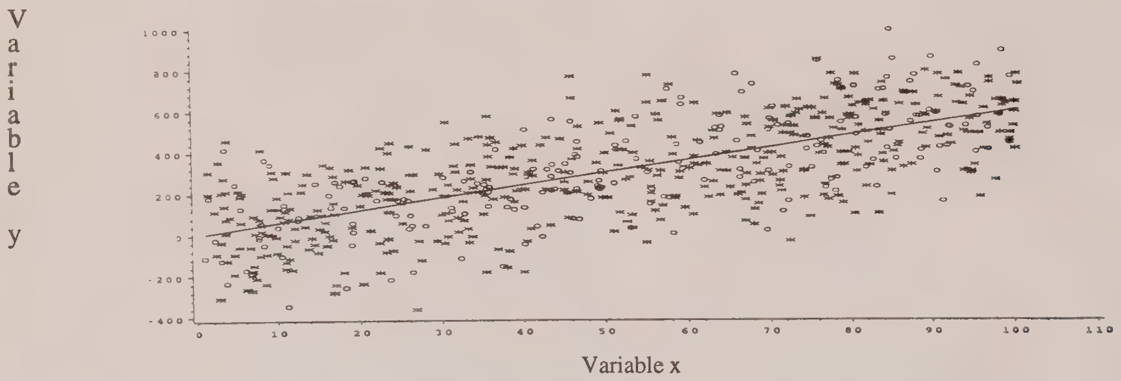


Figure 3a. Deuxième scénario pauvre, 2e itération
Ensemble des appariements faux et 5 % des appariements vrais, données réelles, chevauchement élevé, 650 points, bêta=5,91 R^2 =0,48

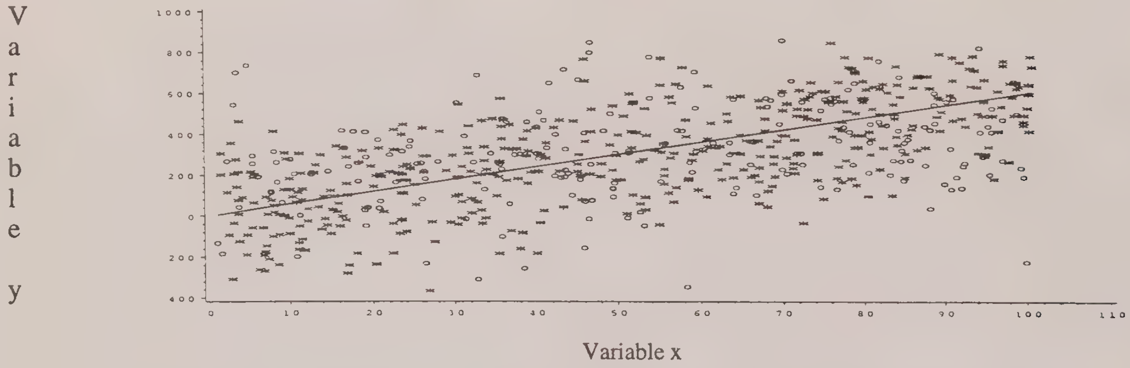


Figure 3b. Deuxième scénario pauvre, 2e itération
Ensemble des appariements faux et 5 % des appariements vrais, données observées, chevauchement élevé, 650 points, bêta=4,75, R^2 =0,33

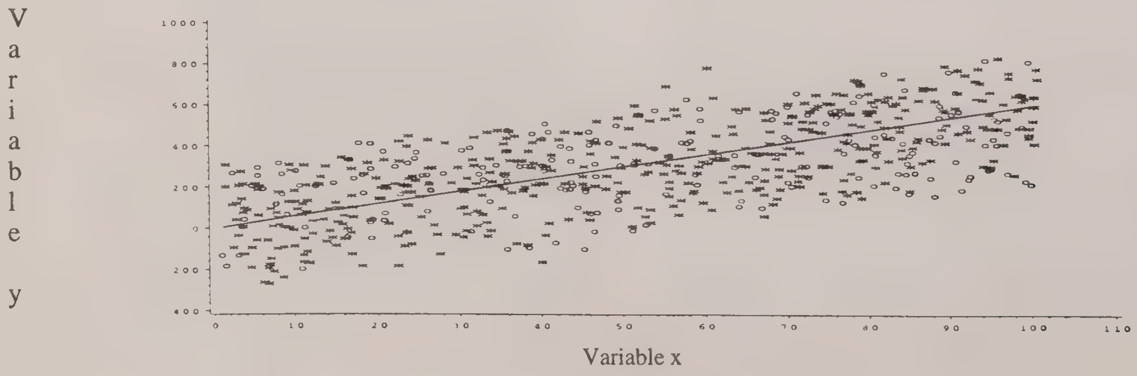


Figure 3c. Deuxième scénario pauvre, 2e itération
Ensemble des appariements faux et 5 % des appariements vrais, valeurs aberrantes-données corrigées, 650 points, bêta=5,26, R^2 =0,47

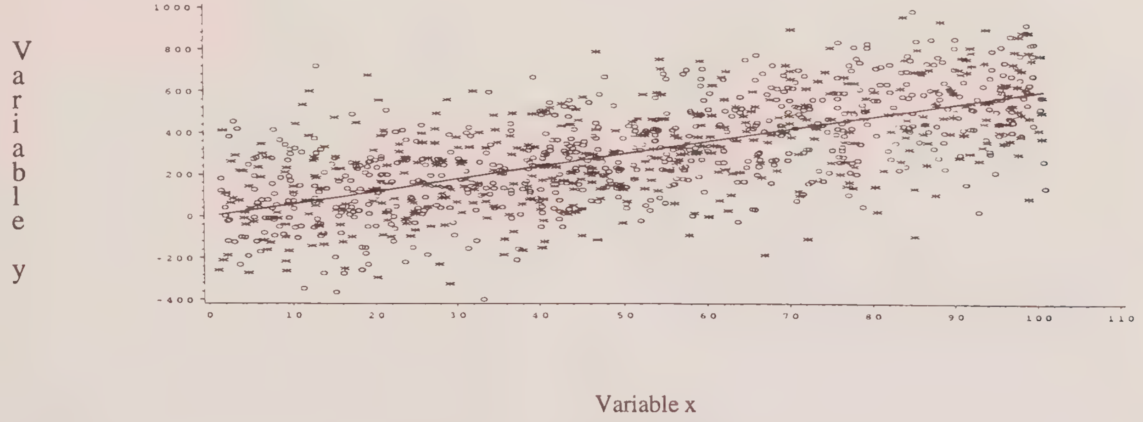


Figure 2a. Deuxième scénario pauvre, 1e itération
Ensemble des appariements faux et 5 % des appariements vrais, données réelles, chevauchement élevé, 1104 points, bêta=5,85, R^2 =0,43

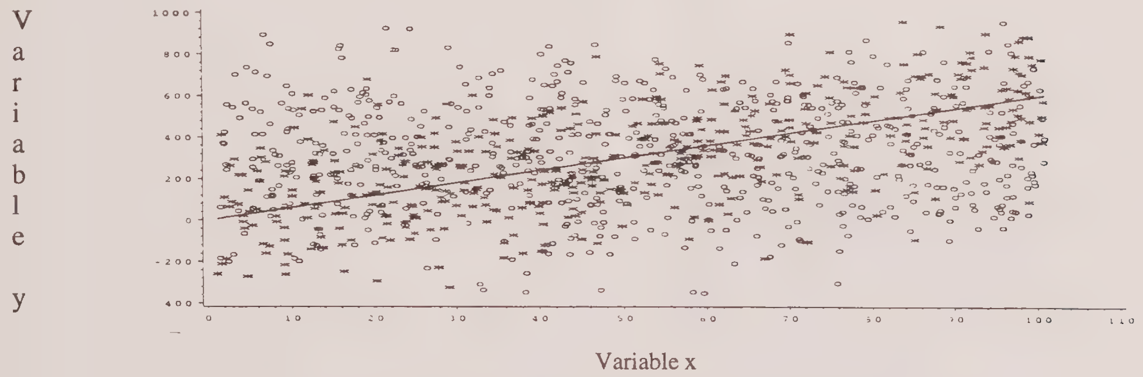


Figure 2b. Deuxième scénario pauvre, 1e itération
Ensemble des appariements faux et 5 % des appariements vrais, données observées, chevauchement élevé, 1104 points, bêta=2,47, R^2 =0,07

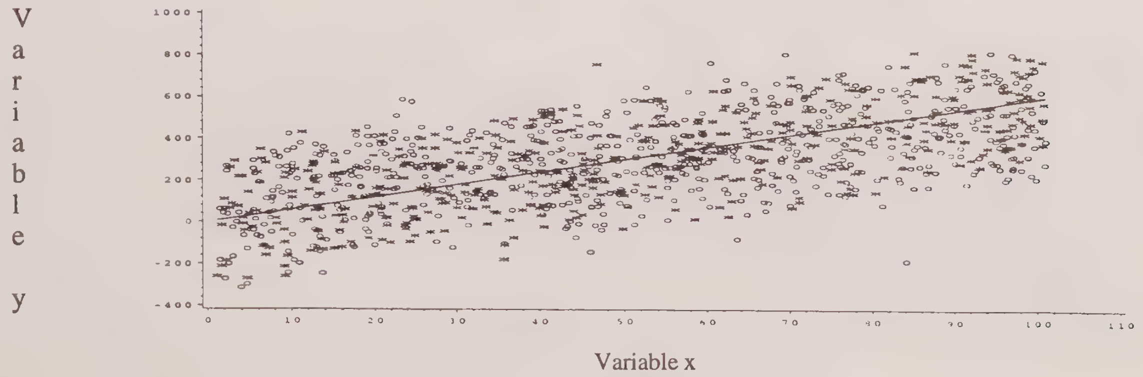


Figure 2c. Deuxième scénario pauvre, 1e itération
Ensemble des appariements faux et 5 % des appariements vrais, valeurs aberrantes-données corrigées, 1104 points, bêta=4,78, R^2 =0,40

considérations analytiques et de notre expérience – sont prises en considération. Pour les paires incluses dans notre analyse, cette restriction fait en sorte que le nombre d'appariements faux dépasse largement le nombre d'appariements vrais. (Rappelons à nouveau que ceci est fait dans le but d'accentuer l'effet visuel des rejets d'appariement et d'accroître la difficulté du problème).

Pour illustrer la situation des données et la technique de modélisation, nous présentons un triplé de tracés. Le premier tracé illustre la situation des données réelles, comme si chaque enregistrement d'un fichier était lié à l'enregistrement auquel il correspond vraiment dans l'autre fichier. Les paires de données quantitatives correspondent au portrait réel. Le deuxième tracé illustre les données observées. On constate qu'une forte proportion de paires comportent des erreurs, car elles correspondent à des appariements faux. Pour obtenir le troisième tracé, nous utilisons un modèle avec un petit nombre de paires (environ 100) dans lesquelles les valeurs aberrantes sont remplacées par des paires où l'on substitue la valeur observée de Y par une valeur prévue de Y.

4.1 Relation de régression vraie initiale

La figure 2a illustre la relation de régression vraie réelle et le nuage de points qui y correspond, pour une de nos simulations, qui seraient obtenus s'il n'y avait pas d'erreurs d'appariement. Dans cette figure et celles qui suivent, la courbe vraie de régression est toujours indiquée pour fins de référence. Enfin, la pente de la population vraie, ou coefficient bêta (à 5,85), et la valeur de R^2 (à 43 %) sont fournies pour les données (échantillon de paires) affichées.

4.2 Régression après l'étape initiale CE→AR

La figure 2b illustre la régression des liens observés réels – non pas des liens que l'on devrait obtenir dans une situation optimale, mais ceux obtenus dans une situation très imparfaite. Fait peu surprenant, nous n'observons qu'une faible relation de régression de Y à X. La pente observée, ou coefficient *bêta*, diffère sensiblement de sa valeur réelle (2,47 contre 5,85). La valeur de R^2 est elle aussi affecée – de 43 %, elle diminue à 7 %.

4.3 Analyse de régression après la première étape combinée CE→AR→VI→AR

La figure 2c complète notre illustration du premier cycle du processus itératif que nous utilisons. Les données dans le graphique illustré ont été corrigées comme suit. Premièrement, en utilisant uniquement les 99 cas pour lesquels le poids d'appariement était supérieur ou égal à 3, nous avons tenté d'améliorer les piètres résultats de la figure 2b. À partir de cet ajustement provisoire, nous avons obtenu des valeurs prévues pour tous les cas appariés; par la suite, les valeurs aberrantes dont le résidu était supérieur ou égal à 460 ont été supprimées et l'analyse de régression a été ajustée de nouveau en fonction des paires restantes. Cette nouvelle équation, utilisée à la figure 2c, est représentée essentiellement par $Y = 4{,}78X + \varepsilon$, avec une variance de 40 000. En utilisant notre approche antérieure (Scheuren et Winkler 1993), un autre ajustement a été fait du coefficient *bêta* estimé, lequel est passé de 4,78 à 5,4. Si une paire d'enregistrements appariés donnait une valeur aberrante, alors les valeurs prévues (non illustrées) obtenues à partir de l'équation $Y = 5{,}4X$ étaient imputées. Si la paire ne donnait pas de valeur aberrante, la valeur observée était utilisée comme valeur prévue.

4.4 Deuxième régression de référence vraie

La figure 3a illustre un nuage de points de X et Y, que l'on obtiendrait s'il s'agissait d'appariements vrais basés sur une deuxième étape de CE. À noter que la série de paires couplées diffère ici quelque peu de la précédente, parce que nous avons utilisé les résultats de la régression pour faciliter le couplage. En termes plus précis, pour la deuxième étape de CE, nous avons utilisé les valeurs prévues de Y tel qu'obtenues précédemment; nous disposions donc de plus d'information sur laquelle baser le couplage. Cela signifie que nous avons obtenu un différent groupe d'enregistrements couplés après la deuxième étape de CE. Comme la qualité du couplage était sensiblement améliorée, il y a eu moins de faux appariements. En conséquence, la taille de notre échantillon formé de tous les faux appariements et de 5 % des appariements vrais a diminué, passant de 1 104 – aux figures 2a à 2c – à 650 aux figures 3a à 3c. Durant cette deuxième itération, la pente vraie ou coefficient bêta et les valeurs de R^2 sont demeurés presque identiques pour ce qui est de la pente estimée (5,85 contre 5,91) et de l'ajustement (43 % contre 48 %).

4.5 Analyse de régression après la deuxième étape CE→AR

À la figure 3b, nous observons une amélioration significative de la relation entre Y et X, à partir des liens observés réels après la deuxième étape de CE. La pente estimée est ainsi passée de 2,47 (initialement) à 4,75. Même si cette valeur demeure trop faible, il y a eu néanmoins nette amélioration. Une amélioration similaire a été observée au niveau de l'ajustement, qui est passé de 7 % à 33 %.

4.6 Analyse de régression après la deuxième étape combinée CE→AR→VI→AR

La figure 3c vient compléter l'illustration du deuxième cycle de notre processus itératif. Les données ont été vérifiées comme suit. À partir de l'ajustement (d'après le paragraphe 4.5), nous avons obtenu une autre série de valeurs prévues pour tous les cas appariés (comme au paragraphe 4.3). La nouvelle équation est représentée essentiellement par $Y = 5{,}26X + \varepsilon$, avec une variance d'environ 35 000. Si une paire d'enregistrements appariés donnait une valeur aberrante, alors les valeurs prévues obtenues à partir de l'équation $Y = 5{,}3X$ étaient imputées. S'il n'y avait pas de valeur aberrante, la valeur observée était utilisée comme valeur prévue.

Figure 1a. Premier scénario d'appariement pauvre

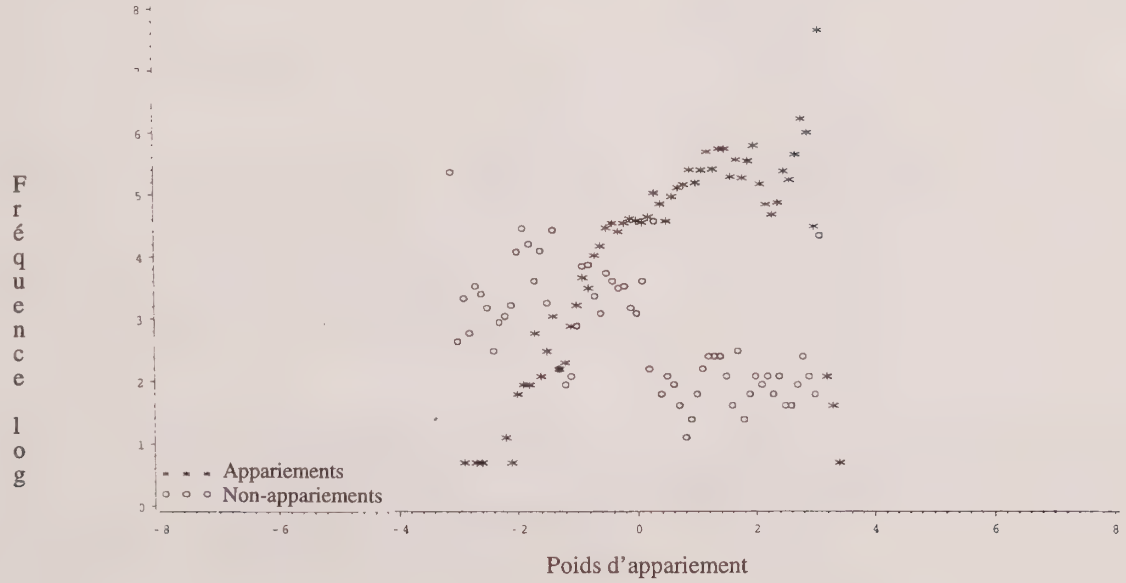


Figure 1b. Deuxième scénario d'appariement pauvre

visé étant d'obtenir des couplages qu'un praticien peu expérimenté pourrait choisir. Cette situation se compare à celle où seraient utilisées des listes administratives de personnes pour lesquelles l'information d'appariement serait de piètre qualité. Le taux de non-appariement vrai a été ici de 10,1 %.

3.3 Deuxième scénario «pauvre» (figure 1b)

Le deuxième scénario d'appariement pauvre consistait à utiliser le nom de famille, le prénom et une variante de l'adresse. Des erreurs typographiques mineures ont été introduites séparément dans le tiers des noms de famille et le tiers des prénoms, dans un des deux fichiers. Des erreurs typographiques graves ont été introduites dans le quart des adresses du même fichier. Les probabilités d'appariement ont été choisies de manière à s'écarter sensiblement du niveau optimal, le but visé étant de représenter une situation qui se produit fréquemment avec des listes d'entreprises pour lesquelles le responsable du couplage a peu d'emprise sur la qualité. Il est souvent très difficile de comparer efficacement l'information sur le nom – une caractéristique d'identification clé – avec les listes d'entreprises. Le taux de non-appariement vrai a été ici de 14,6 %.

3.4 Résumé des scénarios d'appariement

De toute évidence, notre capacité de distinguer les couplages vrais des non-couplages vrais varie sensiblement en fonction du scénario. Dans le premier scénario, le chevauchement – illustré par les courbes de la fréquence (exprimée selon une échelle logarithmique) en fonction du poids – est substantiel (figure 1a); dans le deuxième scénario, le chevauchement des courbes de la fréquence (échelle logarithmique) en fonction du poids est presque total (figure 1b). Lors de travaux antérieurs, nous avons démontré que notre méthode d'ajustement théorique donnait de bons résultats lorsqu'on utilise les taux d'appariement vrais connus dans nos ensembles de données. Dans les cas où les courbes représentant les couplages vrais sont assez bien séparées de celles illustrant les non-couplages vrais, nous avons pu estimer avec exactitude les taux d'erreur par la méthode mise au point par Belin et Rubin (1995), et notre méthode pourrait être utilisée en pratique. Cette méthode de Belin et Rubin n'avait pu fournir d'estimations exactes des taux d'erreur dans le scénario d'appariement pauvre décrit antérieurement (premier scénario pauvre décrit ici), mais notre méthode d'ajustement théorique avait néanmoins donné de bons résultats. C'est ce qui nous a amenés à reconnaître qu'il nous fallait, soit améliorer la méthode de Belin-Rubin, soit élaborer des méthodes faisant un plus grand usage des données disponibles. (Incidemment, cette conclusion découlant de nos travaux antérieurs a mené, après quelques faux départs, à la présente méthode).

3.5 Scénarios quantitatifs

Après avoir précisé les situations de couplage, nous avons utilisé le SAS pour générer des données selon la méthode des moindres carrés ordinaires, d'après le modèle $Y = 6X + \varepsilon$. Les valeurs de X ont été choisies de manière à être distribuées uniformément entre 1 et 101. Les termes d'écart, sont normaux et homoscédastiques, avec des variances respectives de 13 000, 36 000 et 125 000. Les régressions ainsi obtenues de Y en fonction de X ont des valeurs de R^2 dans la population appariée vraie qui correspondent respectivement à 70 %, 47 % et 20 %. Il est difficile de faire un appariement avec des données quantitatives car, pour chaque enregistrement dans un fichier, il existe des centaines d'enregistrements dont les valeurs quantitatives se rapprochent de celles de l'enregistrement qui constitue un appariement vrai. Pour rendre la modélisation et l'analyse encore plus difficiles dans le scénario à chevauchement élevé, nous avons utilisé tous les faux appariements et seulement 5 % des appariements vrais; dans le scénario à chevauchement moyen, nous avons utilisé tous les faux appariements et seulement 25 % des appariements vrais. (Nota: Afin d'accentuer l'effet visuel, nous avons introduit ici une autre étape d'échantillonnage aléatoire, afin que le lecteur puisse mieux «visualiser» dans les figures les effets d'un appariement médiocre. Cet échantillon dépend du statut d'appariement et se limite seulement aux cas appariés – correctement ou incorrectement.)

Une hypothèse pratique essentielle de la présente analyse est que les analystes peuvent produire un modèle raisonnable (estimation subjective) des relations entre les données quantitatives non communes. Pour la modélisation initiale dans l'exemple empirique présenté ici, nous utilisons le sous-ensemble de paires pour lesquelles le poids d'appariement est élevé et le taux d'erreur est faible. Le nombre de faux appariements dans ce sous-ensemble est donc maintenu à un minimum. Même si, ni la méthode de Belin et Rubin (1995), ni celle de Winkler (1994) exigeant une intervention ponctuelle, n'a pu être utilisée pour estimer les taux d'erreur, nous croyons qu'il est possible pour une personne expérimentée dans l'appariement de choisir un ensemble de paires à faible taux d'erreur, même dans le deuxième scénario pauvre.

4. RÉSULTATS DE LA SIMULATION

La majeure partie de cette section est consacrée à la présentation des graphiques et des résultats de l'ensemble du processus appliqué au deuxième scénario pauvre, où la valeur de R^2 est modérée et où l'intersection entre les deux fichiers est grande. Ces résultats sont ceux qui illustrent le mieux les procédures définies dans le présent document. À la fin de la section (paragraphe 4.8), nous résumons les résultats pour l'ensemble des valeurs de R^2 et tous les chevauchements. Afin d'accroître encore davantage la difficulté de la modélisation et d'illustrer la puissance des méthodes de couplage analytique, nous utilisons tous les appariements faux ainsi qu'un échantillon aléatoire formé de seulement 5 % des appariements vrais. Seules les paires dont le poids d'appariement est supérieur à une borne inférieure – laquelle a été déterminée en fonction de

deux fichiers $A \times B$ les paires entre M – l'ensemble des couplages vrais – et U, l'ensemble des non-couplages vrais. À partir de concepts rigoureux introduits par Newcombe (p. ex. Newcombe et coll. 1959; Newcombe, Fair et Lalonde 1992), Fellegi et Sunter (1969) ont examiné les rapports R des probabilités sous la forme

$$R = \Pr((\gamma \in \Gamma \mid M) / \Pr((\gamma \in \Gamma \mid U)$$

où γ est une configuration de concordance arbitraire dans l'espace de comparaison Γ. Γ, par exemple, pourrait être formé de huit configurations représentant une concordance simple (ou non) en regard du nom de famille, du prénom et de l'âge. Ou encore, chaque $\gamma \in \Gamma$ pourrait représenter la fréquence relative à laquelle des noms de famille particuliers, comme Scheuren ou Winkler par exemple, sont présents. Les champs comparés (nom de famille, prénom, âge) sont désignés *variables d'appariement*. La règle de décision est définie comme suit:

Si $R >$ *Limite supérieure*, alors désigner la paire comme un couplage.

Si *Limite inférieure* $\leq R \leq$ *Limite supérieure*, alors désigner la paire comme un couplage possible et la soumettre à une révision manuelle.

Si $R <$ *Limite inférieure*, alors désigner la paire comme un non-couplage.

Fellegi et Sunter (1969) ont démontré que cette règle de décision était optimale car, pour toute paire dont les bornes sont fixes dans R, la région médiane est réduite au minimum en regard de l'ensemble des règles de décision, dans le même espace de comparaison Γ. Les seuils d'inclusion, *Supérieur* et *Inférieur*, sont déterminés par les bornes de l'erreur. Nous désignons le ratio R, ou toute transformation monotone croissante de ce rapport (généralement un logarithme), comme un *poids d'appariement* ou *poids de concordance totale*.

L'introduction de systèmes informatiques peu coûteux a favorisé la prolifération des travaux sur les techniques de couplage d'enregistrements (p. ex. Jaro 1989; Newcombe et coll. 1992; Winkler 1994, 1995). Les nouvelles méthodes informatiques réduisent, et parfois même éliminent, les besoins en révision manuelle lorsque le nom, l'adresse et les autres renseignements servant à l'appariement sont de qualité raisonnable. Le compte rendu d'une récente conférence internationale sur le couplage d'enregistrements vient confirmer ces notions et pourrait constituer en soi la meilleure référence (Alvey et Jamerson 1997).

3. CADRE DE SIMULATION

3.1 Scénarios d'appariement

Aux fins de nos simulations, nous avons utilisé un scénario selon lequel il est pratiquement impossible de distinguer les appariements des non-appariements; lors de travaux antérieurs (Scheuren et Winkler 1993), nous avions examiné trois scénarios dans lesquels les appariements étaient plus faciles à distinguer des non-appariements.

L'idée générale dans ces deux documents demeure toutefois la même, à savoir produire des données ayant des propriétés de distribution connues, attribuer les données aux deux fichiers à apparier, puis évaluer l'effet d'une quantité croissante d'erreurs d'appariement sur les analyses. Comme les méthodes présentées ici donnent de meilleurs résultats que celles proposées antérieurement, nous n'examinons ici qu'un scénario d'appariement qualifié de «deuxième scénario pauvre», car celui-ci est plus difficile que le scénario pauvre (le plus difficile) que nous avions examiné antérieurement.

Nous avons commencé avec deux fichiers de la population (effectif de 12 000 et 15 000), contenant tous deux de bonnes données d'appariement et pour lesquels le véritable statut d'appariement était connu. Les cadres ont été définis comme suit: intersection élevée, moyenne ou faible, selon le nombre de cas dans le petit fichier qui étaient également inclus dans le grand fichier. Dans la première situation (inclusion élevée), environ 10 000 cas sont présents dans les deux fichiers, ce qui donne un taux d'inclusion ou d'intersection par rapport au petit fichier (ou fichier de base) d'environ 83 %. Dans le scénario d'intersection moyenne, nous avons prélevé un échantillon d'un fichier, de manière à ce que l'intersection des deux fichiers à apparier soit d'environ 25 %. Enfin, dans le scénario à faible intersection, les échantillons prélevés des deux fichiers étaient tels que le taux d'intersection entre les fichiers à apparier était d'environ 5 %. De toute évidence, le nombre de cas intersectés limite le nombre d'appariements vrais que l'on peut obtenir.

Nous avons ensuite généré les données quantitatives ayant des propriétés de distribution connues et les avons attribuées aux fichiers. Ces variations sont décrites ci-après et illustrées à la figure 1, où sont représentés le scénario pauvre (désigné «premier scénario pauvre») décrit dans notre article de 1993 et le «deuxième scénario pauvre» retenu pour la présente analyse. Dans cette figure, le poids d'appariement – le logarithme de R – est représenté graphiquement en abscisse en fonction de la fréquence – elle aussi exprimée selon une échelle logarithmique – en ordonnée. Les appariements (ou couplages vrais) sont représentés par un astérisque (*) et les non-appariements (non-couplages vrais) sont représentés par un petit cercle (o).

3.2 Premier scénario «pauvre» (figure 1a)

Le premier scénario d'appariement pauvre consistait à utiliser le nom de famille, le prénom, une variante de l'adresse et l'âge. Des erreurs typographiques mineures ont été introduites séparément dans un cinquième des noms de famille et le tiers des prénoms, dans un des fichiers. Des erreurs typographiques moyennement graves ont été incluses séparément dans le quart des adresses du même fichier. Les probabilités d'appariement ont été choisies de manière à s'écarter sensiblement du niveau optimal, le but

communes. Bien que nous ne puissions le garantir, nous croyons que les méthodes présentées ici donneront assez souvent des résultats concluants, de sorte que l'on peut leur attribuer une valeur générale, à la condition d'avoir un point de départ acceptable.

1.3 Approche fondamentale

Les fondements intuitifs de nos méthodes s'appuient sur les techniques aujourd'hui bien connues du couplage d'enregistrements probabiliste (CE) et de la vérification et imputation (VI). Les principes modernes du CE ont été introduits par Newcombe (Newcombe et coll. 1959) et formalisés mathématiquement par Fellegi et Sunter (1969). Des méthodes récentes sont décrites dans Winkler (1994, 1995). La VI est habituellement utilisée pour éliminer les données erronées des fichiers. Les méthodes les plus pertinentes sont celles basées sur le modèle de VI de Fellegi et Holt (1976).

Pour adapter une analyse statistique en fonction de l'erreur d'appariement, nous utilisons une démarche récursive très puissante, en quatre étapes. Nous commençons par une technique améliorée de CE (p. ex. Winkler 1994; Belin et Rubin 1995), pour définir un sous-ensemble de paires d'enregistrements dans lesquelles on estime que le taux d'erreur d'appariement est très faible. Nous procédons ensuite à une analyse de régression (AR) des enregistrements couplés avec faible taux d'erreur, puis nous corrigeons partiellement le modèle de régression d'après les paires qui restent, en appliquant les méthodes précédentes (Scheuren et Winkler 1993). Nous améliorons ensuite le modèle de VI par les méthodes traditionnelles de détection des valeurs aberrantes, afin de vérifier et d'imputer les valeurs aberrantes dans le reste des paires couplées. Une autre analyse de régression (AR) est faite à ce stade-ci et ces résultats sont intégrés au processus de couplage en vue de l'améliorer. Le cycle se poursuit ainsi jusqu'à ce que les résultats d'analyse désirés cessent de changer. Ces méthodes de *couplage analytique* peuvent être représentées schématiquement par la formule suivante:

$$\begin{array}{c} \nearrow \text{AR} \searrow \\ \text{CE} \leftarrow \text{AR} \leftarrow \text{VI} \end{array}$$

1.4 Aperçu des sections qui suivent

Le présent article se divise en cinq sections, incluant l'introduction. Dans la deuxième section, nous faisons un bref examen des méthodes de vérification et imputation (VI) et de couplage d'enregistrements (CE). Notre but n'est pas de décrire ces méthodes en détail, mais plutôt d'en préciser le cadre aux fins de la présente application. L'analyse de régression (AR) étant une technique bien connue, nous ne l'aborderons qu'en rapport avec les simulations particulières examinées (section 3). Ces simulations ont pour but de présenter des scénarios d'appariement plus difficiles que ceux qui sont habituellement traités par la plupart des responsables du couplage. Nous utilisons des données quantitatives qui sont à la fois faciles à comprendre et difficiles à utiliser pour l'appariement; les résultats obtenus sont présentés à la quatrième section. L'article se termine, à la section cinq, par un énoncé de quelques conclusions et de domaines d'études futurs.

2. MÉTHODES DE VI ET DE CE

2.1 Vérification et imputation

Les méthodes de vérification des microdonnées avaient habituellement pour but d'éliminer les incohérences logiques dans les bases de données. Le logiciel était construit selon des règles de type «*si-alors*», qui étaient spécifiques de la base de données et très difficiles à mettre à jour ou à modifier pour les garder actuelles. Les méthodes d'imputation faisaient partie de la série de règles *si-alors* mais pouvaient donner lieu malgré tout au rejet des enregistrements révisés, au moment de la vérification. À la suite d'une percée théorique importante, qui est venue rompre avec les méthodes statistiques jusque là utilisées, Fellegi et Holt (1976) ont proposé des méthodes basées sur la recherche opérationnelle, qui permettent à la fois de vérifier la cohérence logique d'un système de vérification et de toujours pouvoir mettre à jour un enregistrement rejeté à la vérification à partir de valeurs imputées. De cette manière, l'enregistrement révisé satisfait à toutes les vérifications. Autre avantage du système Fellegi et Holt (1976) celui-ci permet de lier directement la méthode de vérification aux méthodes actuelles d'imputation des microdonnées (p. ex. Little et Rubin 1987).

Bien que le présent article porte uniquement sur les données continues, les techniques de VI peuvent également s'appliquer aux données discontinues ou à une combinaison de données continues et discontinues. Aux fins du présent exemple, supposons que nous avons des données continues où l'ensemble des vérifications pourrait consister en des règles pour chaque enregistrement, ayant la forme suivante:

$$c_1X < Y < c_2X$$

En termes plus précis,

> On peut s'attendre à ce que Y soit supérieur à c_1X et inférieur à c_2X; par conséquent, si Y est inférieur à c_1X et supérieur à c_2X, alors l'enregistrement de données devrait être révisé (à partir des ressources et autres considérations pratiques déterminant les bornes effectives utilisées).

Dans l'exemple présenté, Y pourrait représenter le salaire total; X être le nombre d'employés et c_1 et c_2 être des constantes où $c_1 < c_2$. Lorsqu'une paire (X, Y) associée à un enregistrement est rejetée à la vérification, nous pouvons remplacer, disons Y, par une estimation (ou prévision).

2.2 Couplage d'enregistrements

Le processus de couplage d'enregistrements consiste à répartir, à l'intérieur d'un espace provenant du produit de

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 171-180
Statistique Canada

Analyse de régression des fichiers de données appariés par ordinateur - Partie II

FRITZ SCHEUREN et WILLIAM E. WINKLER[1]

RÉSUMÉ

Dans bien des cas, les meilleures décisions en matière de politiques sont celles qui peuvent s'appuyer sur des données statistiques, elles-mêmes obtenues d'analyses de microdonnées pertinentes. Cependant, il arrive parfois que l'on dispose de toutes les données nécessaires mais que celles-ci soient réparties entre de multiples fichiers pour lesquels il n'existe pas d'identificateurs communs (p. ex. numéro d'assurance sociale, numéro d'identification de l'employeur ou numéro de sécurité sociale). Nous proposons ici une méthode pour analyser deux fichiers de ce genre: 1) lorsqu'il existe des informations communes non uniques, sujettes à de nombreuses erreurs et 2) lorsque chaque fichier de base contient des données quantitatives non communes qui peuvent être reliées au moyen de modèles appropriés. Une telle situation peut se produire lorsqu'on utilise des fichiers d'entreprises qui n'ont en commun que l'information – difficile à utiliser – sur le nom et l'adresse, par exemple un premier fichier portant sur les produits énergétiques consommés par les entreprises et l'autre fichier regroupant les données sur le type et la quantité de biens produits. Une autre situation similaire peut survenir avec des fichiers sur des particuliers, dont le premier contiendrait les données sur les gains, le deuxième, des renseignements sur les dépenses reliées à la santé et le troisième, des données sur les revenus complémentaires. Le but de la méthode présentée est de réaliser des analyses statistiques valables, avec production ou non de fichiers de microdonnées pertinentes.

MOTS CLÉS: Vérification; imputation; couplage d'enregistrements; analyse de régression.

1. INTRODUCTION

1.1 Cadre d'application

Pour modéliser adéquatement le rendement énergétique, un économiste peut avoir besoin de microdonnées propres à l'entreprise sur sa consommation de carburant et de matières premières – lesquelles données ne sont disponibles qu'auprès de l'organisme A – et des microdonnées correspondantes sur les biens produits par l'entreprise, lesquelles microdonnées sont disponibles uniquement de l'organisme B. Autre exemple, pour établir un modèle sur la santé des personnes vivant dans la société, le démographe ou le responsable de l'élaboration des politiques en matière de santé peut avoir besoin de données propres à la personne, par exemple l'information sur les personnes touchant des prestations sociales des organismes B1, B2 et B3, l'information correspondante sur le revenu, obtenue de l'organisme I et l'information sur les services de santé, fournie par les organismes H1 et H2. Or une telle modélisation n'est possible que si l'analyste a accès aux microdonnées et s'il existe des identificateurs communs et uniques (p. ex. Oh et Scheuren 1975; Jabine et Scheuren 1986). Cependant, si les seuls identificateurs communs qui existent sont sujets à erreurs ou qu'ils ne sont pas uniques – ou les deux – alors il faut utiliser une technique d'appariement probabiliste (p. ex. Newcombe, Kennedy, Axford et James 1959; Fellegi et Sunter 1969).

1.2 Liens avec des travaux antérieurs

Dans le cadre de travaux antérieurs (Scheuren et Winkler 1993), nous avions proposé une théorie qui permettait de corriger avec justesse les analyses de régression élémentaires en fonction de l'erreur d'appariement, à partir des données sur la qualité de l'appariement. Pour ces travaux, nous nous étions basés largement sur la technique d'estimation du taux d'erreur de Belin et Rubin (1995). D'autres travaux effectués par la suite (Winkler et Scheuren 1995, 1996) ont démontré qu'il était possible d'améliorer encore davantage cette technique en utilisant des données quantitatives non communes provenant des deux fichiers, de manière à améliorer l'appariement et à corriger les analyses statistiques en fonction de l'erreur d'appariement. La principale exigence – même dans les cas qui semblaient jusque là impossibles – était qu'il devait exister un modèle raisonnable des relations entre les données quantitatives non communes. Dans l'exemple empirique présenté ici, nous utilisons des données pour lesquelles un très petit sous-ensemble de paires peut être apparié de façon exacte, à partir uniquement de l'information sur le nom et l'adresse, là où il existe une corrélation tout au moins modérée entre les données quantitatives non communes. Dans d'autres cas, les chercheurs pourraient utiliser un petit fichier de microdonnées qui représente exactement les relations entre des données non communes pour un ensemble de gros fichiers administratifs ou s'appuyer uniquement sur une présomption raisonnable des liens entre les données non

[1] Fritz Scheuren, Ernst and Young, 1225 Connecticut Avenue, N.W., Washington, DC 20036, U.S.A., Scheuren@aol.com; William E. Winkler, U.S. Bureau of the Census, Washington, DC 20023, U.S.A.

HALE, A., et MICHAUD, S. (1995). Dependent Interviewing: Impact on Recall and on Labour Market Transitions. Enquête sur la dynamique du travail et du revenu, documents de recherche, 95-06. Statistique Canada.

HIEMSTRA, D., LAVIGNE, M., et WEBBER, M. (1993). Labour Force Classification in SLID: Evaluation of Test 3A Results. Enquête sur la dynamique du travail et du revenu, documents de recherche, 93-14. Statistique Canada.

KAUSHAL, R., et LANIEL, N. (1995). Computer-assisted interviewing data quality test. *Proceedings of the 1993 Annual Research Conference.* U.S. Bureau of the Census, 513-524.

LAVIGNE, M., et MICHAUD, S. (1995). Aspects généraux de l'Enquête sur la dynamique du travail et du revenu. *Recueil des textes des présentations du colloque sur les applications de la statistique*. L'association canadienne française pour l'avancement des sciences.

LYBERG, L., BIEMER, P., COLLINS, M., de LEEUW, E., DIPPO, C., SCHWARZ, N., et TREWIN, D. (1997). *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

MICHAUD, S., LE PETIT, C., et LAVIGNE, M. (1993). Aspects qualitatifs de la collecte du test 3A de l'Enquête sur la dynamique du travail et du revenu, documents de recherche, 93-07. Statistique Canada.

MICHAUD, S., LAVIGNE, M., et POTTLE, J. (1993). Aspects qualitatifs de la collecte du test 3B de l'Enquête sur la dynamique du travail et du revenu, documents de recherche, 93-11. Statistique Canada.

MURRAY T.S., MICHAUD, S., EGAN, M., et LEMAÎTRE, G. (1990). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1990 Annual Research Conference.* U.S. Bureau of the Census.

NICHOLLS II, W.L., et GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part I. *Journal of Official Statistics*, 2, 93-115.

SIMARD, M., et DUFOUR, J. (1995). Impact de l'implantation des interviews assistées par ordinateur comme nouvelle méthode de collecte à l'enquête sur la population active. Rapport technique, division des méthodes d'enquêtes-ménages, Statistique Canada.

SIMARD, M., DUFOUR, J., et MAYDA, F. (1995). The first year of computer-assisted interviewing as the Canadian Labour Force Survey data collection method. *Proceedings of Section on Survey Research Methods, American Statistical Association*, 533-538.

SINGH, M.P., GAMBINO, J., et LANIEL, N. (1993). Research studies for the Labour Force Survey sample redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

STATISTIQUE CANADA (1998). *Méthodologie de l'enquête sur la population active du Canada.* 71-526 au catalogue. À paraître.

TAMBAY, J.-L., et CATLIN, G. (1995). Plan d'échantillonnage de l'Enquête nationale sur la santé de la population, Rapports sur la santé. Catalogue 82-003, Statistique Canada, 7, 31-42.

WILLIAMS, B., et SPAULL, M. (1992). Computer-assisted Personal Interviewing LFS Datellite Test 0691-1191. Rapport interne. Conférence des gestionnaires de ISS, Statistique Canada.

texte affichée, aux fonctions clés préprogrammées et à la facilité de déplacement d'un écran à un autre. De plus, comme on demande aux intervieweurs de travailler sur plus d'une enquête, il faudrait, dans la mesure du possible, faire un effort d'uniformisation des formats d'écran.

En ce qui concerne les composantes matérielles et logicielles, les équipes de travail s'affairent actuellement à choisir la meilleure combinaison. À l'heure actuelle, on utilise différents logiciels pour différentes composantes dans le cadre de plusieurs enquêtes. Afin de normaliser le plus possible les applications disponibles, on projette d'utiliser une plate-forme uniformisée pour toutes les enquêtes dans un environnement Windows. L'environnement Windows devrait donner aux intervieweurs et aux programmeurs une plus grande souplesse. Il faut aussi repenser les systèmes de sécurité, pour les rendre conforme à la technologie adoptée et pour satisfaire aux exigences de Statistique Canada. Il faut tenter d'harmoniser les questions d'une enquête à l'autre, ce qui permettrait de modulariser davantage la programmation de l'IAO. Le fardeau du répondant en serait lui aussi allégé.

Le nouveau système devra pouvoir tenir compte des exigences tant passées que présentes. Par exemple, les caractéristiques des systèmes sont réexaminées sur la base des rapports d'étapes fournis au personnel opérationnel pour déterminer quels sont les points à améliorer. Comme on l'a noté dans la section 4, un certain nombre d'autres possibilité sont envisagées, telles que la formation interactive des intervieweurs, des modules de formation spéciaux, la possibilité de mener des réinterviews et de meilleurs instruments de dépistage. Grâce à ces fonctions, on pourra mieux tirer parti de la souplesse acquise par l'automatisation du processus.

On est également en train de concevoir un nouveau système de gestion des cas. L'un des impératifs visés est d'installer un système de communication robuste qui permettra la transmission uniforme des changements et une fonction de réplication. On espére pouvoir élaborer un système informatique qui sera utilisé pendant de nombreuses années à venir, mais la réalité actuelle semble suggérer que l'IAO devrait continuer d'évoluer rapidement. Eu égard à cette rapide évolution technologique (on n'a qu'à penser à Internet), le défi présent consiste à mettre au point un système souple qui pourra être facilement adapté sans nécessiter une restructuration complète.

REMERCIEMENTS

Les auteurs veulent remercier les nombreuses personnes de la Division des méthodes d'enquêtes des ménages, de la Division des méthodes d'enquêtes sociales, de la Division des enquêtes-ménages et de la Division des opérations d'enquêtes qui, au fil des années, ont contribué à l'élaboration de l'IAO à Statistique Canada. C'est grâce à leur travail que le présent document a été rendu possible. Nous voulons également remercier Ann Brown, Brian Williams, Jean-Louis Tambay et Frank Mayda de leurs précieux commentaires qui ont permis d'améliorer la qualité de ce document.

BIBLIOGRAPHIE

ALLARD, B., BRISEBOIS, F., DUFOUR, J., et SIMARD, M. (1996). How do interviewers do their job? A look at new data quality measures for the Canadian Labour Force Survey. Présenté à l'International Conference on Computer-assisted Survey Information Collection.

ALLARD, B., DUFOUR, J., SIMARD, M., et BASTIEN, J.-F. (1996). Pourquoi refuse-t-on de participer aux enquêtes? Le cas de l'Enquête sur la population active. Direction de la méthodologie, document de travail, DMEM, 96-003F. Statistique Canada.

BRISEBOIS, F., DUFOUR, J., et LÉVESQUE, I. (1997). New LFS quality measures. Direction de la méthodologie, document de travail, Statistique Canada. À paraître.

BRODEUR, M., MONTIGNY, G., et BÉRARD, H. (1995). Challenge in developing the National Longitudinal Survey of Children. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-28.

BROWN, A., HALE, A., et MICHAUD, S. (1997). Use of Computer-assisted Interviewing in Longitudinal Surveys. Présenté à l'International Conference on Computer-assisted Survey Information Collection.

CATLIN, G., et INGRAM, S. (1988). The effects of CATI on cost and data quality. Dans *Telephone Survey Methodology*, édité par R.M. Groves et coll., New York: John Wiley and Sons.

CATLIN, G., ROBERTS, K. et INGRAM S. (1996). Validité de l'auto-déclaration des problèmes de santé chroniques lors de l'enquête nationale sur la santé de la population. Présenté au Symposium 96, Erreurs non dues à l'échantillonnage, Statistique Canada.

CLARK, C., MARTIN, J., et BATES, N. (1997). Development and Implementation of CASIC in Government Statistical Agencies. Présenté à l'International Conference on Computer-assisted Survey Information Collection.

DIBBS, R., HALE, A., LOVEROCK, R., et MICHAUD, S. (1995). Some Effects of Computer-assisted Interviewing on the Data Quality of the Survey of Labour and Income Dynamics. Enquête sur la dynamique du travail et du revenu, documents de recherche, 95-07. Statistique Canada.

DREW, D., GAMBINO, J., AKYEAMPONG, E., et WILLIAMS, B. (1991). Plans for the 1991 redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

DUFOUR, J., KAUSHAL, R., CLARK, C., et BENCH, J. (1995). Converting the Labour Force Survey to Computer-assisted Interviewing. Direction de la méthodologie, document de travail, DMEM, 95-009E. Statistique Canada.

DUFOUR, J., SIMARD, M., et MAYDA, F. (1995). The First Year of Computer-assisted Interviewing for the Canadian Labour Force Survey: An Update. Direction de la méthodologie, document de travail, DMEM, 95-011E. Statistique Canada.

portatifs, de sorte que seul l'intervieweur a accès aux renseignements. De plus, les données sont cryptées lorsqu'elles résident dans l'ordinateur portatif.

Les difficultés que pose l'IAO en matière de confidentialité sont très différentes de celles que l'on avait avec la méthode du papier et du crayon. Dans le cas de l'EDTR, les interviews dépendantes posent des difficultés de cet ordre. L'information sur une famille provenant d'une collecte précédente peut devenir délicate dans le cas où, par exemple, le ménage se sépare. Par conséquent, si la nouvelle méthode permet de conduire des interviews dépendantes, celles-ci posent des inconvénients qui doivent être évalués pour chaque situation.

Avec l'apparition de l'auto-interview audio assistée par ordinateur (AIAO-A), il est plus facile de traiter les sujets délicats. Le répondant est relié à l'ordinateur par un casque d'écoute et les questions sont lues par une voix numérisée. Le répondant peut donc choisir s'il veut ou non que les questions s'affichent à l'écran. Grâce à cette technique, le répondant peut remplir le questionnaire de façon parfaitement anonyme. On prévoit commencer à utiliser cet outil dans le cadre de l'ELNEJ avant la fin de l'an 2000.

4.6 Programmes de réinterview

En ce qui concerne les programmes de réinterview, l'IAO offre certains avantages par rapport à l'IPC. Premièrement, la rapidité de la transmission électronique des données réduit les écarts attribuables à des problèmes de mémoire, puisque les réinterviews peuvent avoir lieu dans un délai plus court suivant la première interview. L'observation rigoureuse des règles de réconciliation intégrées dans le logiciel permet d'obtenir une estimation plus précise des erreurs de mesure. Les intervieweurs feuilletent le questionnaire avant de commencer la réinterview. De même, les réconciliations peuvent être faites après un sous-ensemble de questions, à la fin d'une section ou à la fin du questionnaire, et autant de fois qu'il le faut. Les cas de réinterviews sont facilement automatisés et intégrés dans un processus de contrôle de la qualité tenant compte des caractéristiques de l'intervieweur et de l'interview (cas particuliers se rapportant à des problèmes de formation, cas appartenant à un groupe particulier). La qualité des données est meilleure parce qu'un grand nombre de règles de vérification, identiques à celles qui sont appliquées au cours de l'interview sont programmées pour les réinterviews. Les fonctions offertes par le SGC sont également un atout pour le programme de réinterviews: progression du programme de réinterviews, performance et progression des réinterviews, transfert facile des cas, *etc.*

4.7 Formation de l'intervieweur

Avec l'adoption de l'IAO, les intervieweurs ont vu leurs méthodes de travail changer considérablement. La formation s'est révélée une étape essentielle, leur permettant de s'adapter efficacement à cette méthode informatisée de collecte de données. Ils se sont familiarisés à de nouveaux outils de travail (clavier, ordinateur portatif et toutes les procédures informatiques qu'il faut suivre, comme l'enregistrement des données, le chargement des piles et la transmission par modem). Ils ont également dû adapter leur style d'interview aux exigences de l'IAO. Par ailleurs, les nouveaux intervieweurs ont dû se familiariser avec les concepts propres aux enquêtes, les techniques d'interview et l'instrument de collecte. Pour relever ce défi, Statistique Canada a élaboré une stratégie de formation fondée sur l'expérience qu'elle a acquise au cours des essais antérieurs et sur l'expérience de collègues britanniques et américains.

La formation des intervieweurs demeurera l'un des facteurs clés du succès des enquêtes de Statistique Canada, et l'organisme innove constamment dans ce domaine. Par exemple, l'une des initiatives dans le cadre de l'EPA est la mise en application d'une stratégie consistant à permettre aux intervieweurs principaux de recevoir régulièrement une petite tâche d'IAO (environ 15 cas), de sorte qu'ils puissent s'exercer à cette méthode de collecte et se tenir au fait de l'évolution de l'application d'IAO. Outre les cas de formation ordinaires qui sont toujours accessibles dans l'ordinateur, le système IAO offrira aux intervieweurs des modules intégrés au système de collecte et traitant de sujets complexes comme la couverture et les logements multiples, de sorte qu'ils pourront se tenir à jour et réviser différents concepts difficiles.

5. L'AVENIR DE L'IAO À STATISTIQUE CANADA

Dans le nouvel environnement de ressources limitées et de lourd fardeau des répondants, la collecte statistique devient de plus en plus adaptée à chaque enquête. Alors que les enquêtes auprès des entreprises ont pris cette forme depuis un certain temps déjà, la collecte mixte commence à être en demande pour les enquêtes-ménages. La collection centralisée à l'extérieur de la période de collecte pour un nombre limité de répondants peut permettre d'améliorer le taux de réponse (en mettant l'accent sur le dépistage par exemple). L'environnement nécessaire à ce type de collecte ressemble davantage à l'ITAO qui permet la mise en commun de fonctions de bases de données pour un petit échantillon, ainsi que des fonctions de planification d'appels.

On prévoit que, d'ici la fin du siècle, l'application d'IAO et le système de gestion des cas seront complètement repensés. Au cours de ce remaniement, les équipes de travail devront tenir compte non seulement des capacités de l'ordinateur, mais aussi de facteur humain. Ce dernier facteur est important parce que la collecte de données et la qualité des données en dépendent. Les intervieweurs doivent lire à l'écran et faire la saisie des réponses, des tâches qui requièrent des habiletés perceptives et motrices différentes de celles qu'ils utilisaient avec la méthode du papier et crayon. Le libellé des questions est également plus difficile à lire à l'écran, et les intervieweurs disent qu'il est plus ardu de visualiser la structure d'ensemble d'un questionnaire. Il faut donc porter une attention spéciale au design de l'écran, au choix des couleurs, à la quantité de

les intervieweurs, leurs superviseurs et les bureaux régionaux. Depuis que l'IAO a été adoptée pour la première fois, le processus de communication a été sensiblement amélioré, de sorte que chaque intervieweur puisse recevoir ses tâches, la dernière version de l'application ou différents changements. Néanmoins, ce processus doit faire l'objet d'un contrôle permanent. Par exemple, à la fin de la période de collecte, les cas doivent être transmis et supprimés de l'ordinateur de l'intervieweur. La plupart du temps, les cas non transmis sont essentiellement des non-réponses. Comme ces cas ne sont pas transmis au siège social après la fin de la période de collecte, on perd parfois l'information sur les motifs de ces non-réponses. Bien que beaucoup de ces problèmes puissent être repérés durant les essais, il reste qu'il demeure toujours quelques cas exceptionnels.

4.2 Procédés de contrôle pour l'IAO

Le SGC et les applications d'enquêtes ont la capacité de produire de nombreuses bases de données. La quantité de données est souvent écrasante et l'on n'exploite pas réellement ces données à leur potentiel maximal. En outre, la vitesse inhérente à l'IAO fait que l'on n'a pas assez de temps et de ressources pour analyser et contrôler cette masse d'information. Pour le moment, cette information est utilisée après coup, mais il serait grandement souhaitable que l'on puisse l'utiliser pendant que l'enquête est en cours.

Les intervieweurs devraient pouvoir accéder à cette information dans un format intégré. Cependant, il faut un juste équilibre pour éviter l'excès de surveillance qui amènerait les intervieweurs à porter davantage d'attention aux indicateurs de qualité qu'à la qualité des données comme telles. Idéalement, on pourrait analyser plusieurs enquêtes pour relever les problèmes particuliers, et concevoir ensuite des trousses de formation brèves et pertinentes. De plus, les taux de réponse et les taux de couverture pourraient être intégrés pour les enquêtes. Tous ces renseignements pourraient servir à améliorer la gestion du temps ou à préparer de la formation sur des compétences d'interview particulières.

4.3 Vérification en cours de collecte

Bien que l'IAO permette d'inclure un grand nombre de règles de vérification pouvant servir au moment de l'interview, il est important ici de maintenir un équilibre entre les règles programmées dans l'outil de collecte et les règles appliquées au cours du traitement par lots au siège social. Les règles programmées dans l'application prolongent l'interview, ce qui augmente les coûts et le fardeau des répondants. Avec l'évolution technologique rapide que nous devrions connaître d'ici quelques temps, il devrait être possible d'appliquer un plus grand nombre de règles de vérification au cours de l'interview, sans en perturber le rythme. Par ailleurs, toute clarification donnée pendant l'interview améliore la qualité des données. Les données de l'Enquête nationale sur la santé de la population sont de meilleure qualité à la collecte du deuxième trimestre parce que l'on utilise les renseignements du premier trimestre pour alimenter le système de vérification. Par exemple, donner des clarifications au répondant au cours de l'interview nous a permis de découvrir que, dans le cas de la variable arthrite, sur les 7 % de répondants qui indiquent un changement dans leur état entre les deux trimestres, seulement 3,3 % avait réellement connu un changement, alors que pour 3,5 % il s'agissait d'erreurs. Pour de plus amples détails, voir Catlin, Roberts et Ingram (1996).

Avec l'IAO, il est également possible de stocker l'information pour indiquer quelles règles de vérification ont été déclenchées et quelles corrections ont été apportées. Une étude portant sur les règles de vérification les plus souvent déclenchées permettrait de déterminer quelles règles influencent le plus la qualité des données. Une telle étude servirait non seulement à titre informatif, mais ses résultats permettraient de modifier des règles trop strictes et serviraient de base à un système de correction dynamique. Un autre aspect aussi important concerne la facilité avec laquelle l'intervieweur peut faire les corrections nécessaires. S'il suffit de corriger la réponse actuelle ou la réponse précédente à une question, l'intervieweur peut le faire facilement. Par contre, s'il faut vérifier une série de réponses, remonter d'une réponse à l'autre et déterminer laquelle a besoin d'être corrigée, cela peut être trop complexe pour que cette vérification ait lieu durant l'interview.

Outre les problèmes techniques, il existe des problèmes méthodologiques associés à l'incidence des règles de vérification sur la qualité des données. À quelle étape l'application des différentes règles de vérification est-elle la plus efficace? Les règles touchant l'enchaînement du questionnaire et celles qui déterminent quelles personnes sont hors du champ d'application de l'enquête sont essentielles. Les variables clés servant à la stratification a posteriori et aux estimations clés se définissent mieux au moment de l'interview. Le nombre de règles de validation pouvant être intégrées à l'IAO est fonction de la vitesse de l'ordinateur portatif. En outre, lorsque certaines règles sont élaborées pour l'instrument tandis que d'autres sont destinées au traitement central, il faut s'assurer que les deux types de règles n'entrent pas en contradiction.

4.5 Confidentialité des données

La préservation de la confidentialité des données, conformément aux stipulations de la *Loi sur la statistique,* est une des exigences fondamentales qui régissent l'utilisation de l'IAO et des systèmes qui la soutiennent. Pour répondre à cette exigence, on a élaboré un certain nombre de procédures et on a mis en place, notamment, un environnement informatique comportant deux réseaux de communication, un interne et un externe. Les données sont transférées physiquement, sur bande, du réseau externe au réseau interne confidentiel, parce qu'il n'y a pas de connexion entre ces deux réseaux. Il est impossible d'accéder au réseau interne à l'aide d'un modem public. On assure aussi la confidentialité de l'information par le cryptage des données dès que celles-ci doivent être transmises par le réseau téléphonique. De plus, un système de contrôle des accès est intégré dans tous les ordinateurs

3.3 Nouveaux indicateurs de qualité

La méthode IAO adoptée par Statistique Canada pour ses enquêtes-ménages offre un système complexe de contrôle des opérations d'enquête au cours des périodes de collecte pour veiller à ce que tout fonctionne bien. Ce système appelé «système de gestion des cas» (SGC) est un système perfectionné qui permet de gérer toutes les opérations du début à la fin du cycle d'enquête. Ce système est souple, puisqu'il peut être adapté aux besoins des différentes enquêtes-ménages qui l'utilisent. Le SGC exécute trois fonctions principales: i) le cheminement des cas, ii) la production de rapports sur les opérations et iii) l'aide aux intervieweurs. Le module de cheminement dirige les mouvements de cas durant l'enquête, que ce soit de l'intervieweur au bureau régional, du bureau régional au siège social, *etc*. Le deuxième module du SGC produit différents rapports décrivant l'état de l'enquête à un point donné dans le temps, évaluant les performances et le progrès de l'enquête et indiquant l'état des interviews. Toute une gamme de renseignements sont produits par cette deuxième composante du SGC. Enfin, le troisième module permet aux intervieweurs de remplir leurs tâches plus efficacement, au moyen d'options de prises de rendez-vous, d'enregistrement de notes, *etc*.

Par conséquent, ce système offre une masse d'information sur ce qui arrive effectivement sur le terrain au cours d'une enquête; toute mesure prise relativement à un cas est enregistrée par le SGC. Le grand défi dans ce type de système est d'éviter de se perdre dans la grande masse de renseignements disponibles. On a mis sur pied des équipes de travail pour maîtriser ces sources d'information, élaborer de nouveaux indicateurs de qualité en utilisant cette information ou en la combinant avec d'autres renseignements déjà disponibles, trouver des utilisations (formations additionnelles, amélioration de l'instrument de collecte de données) et trouver des manières de présenter ces indicateurs de façon efficace.

On a produit un grand nombre d'indicateurs de qualité (voir Simard et coll. 1995; Allard, Brisebois, Dufour et Simard 1996) à un rythme régulier et à différents niveaux d'intérêt (géographique, intervieweurs, administration). On peut grouper ces indicateurs en deux catégories: information et contrôle. Parmi les indicateurs d'information mentionnons: le nombre de tentatives avant de compléter un cas, la distribution des interviews terminées par jour de collecte, la meilleure combinaison jour-heure pour joindre un répondant, la durée médiane des interviews et le nombre de règles de validation déclenchées et ignorées ou déclenchées et sur lesquelles on a pris des mesures (voir Brisebois, Dufour et Lévesque 1997). Les indicateurs d'information servent à améliorer ou à modifier la stratégie ou le processus de collecte.

En matière de contrôle, on se sert d'une série d'indicateurs pour retracer les irrégularités commises sur le terrain, qu'elles soient humaines ou techniques. Parmi ces indicateurs, on peut mentionner: les appels ou les visites effectués après la date de transmission mais avant la semaine d'enquête, les appels ou les visites faits après le dimanche de la semaine de d'enquête, les périodes de travail trop tôt, les périodes de travail trop tardives, les interviews trop courtes, *etc*. Ces renseignements servent à vérifier si les instructions formulées par le siège social sont suivies et si certains intervieweurs ont besoin de davantage de formation. Toutefois, toutes ces données doivent être analysées avec prudence pour déterminer la cause de l'irrégularité. Par exemple, une interview menée à 4 h 30 du matin peut très bien l'avoir été à la demande du répondant, un fermier par exemple, à moins que l'horloge de l'ordinateur ne soit mal réglée (voir Brisebois et coll. 1997).

L'IAO permet également aux intervieweurs d'inclure un commentaire pour chaque question ou d'expliquer pourquoi tel code a été donné. Il est donc possible d'adapter la formation en fonction de ces commentaires, de mieux comprendre les enquêtes et, par conséquent, de mieux les adapter aux réalités du terrain. Par exemple, cette fonction a permis de mener une étude spéciale sur les motifs de refus de participer à l'une des enquêtes-ménages de Statistique Canada. Une telle étude aurait auparavant nécessité beaucoup d'efforts (voir Allard, Dufour, Simard et Bastien 1996).

4. LES DÉFIS ACTUELS DE L'IAO

Cette section décrit les défis à long terme qui se posent en matière d'élaboration, de mise en oeuvre et de compréhension de l'utilisation de l'IAO pour les applications d'enquêtes. Les puissants outils rendus accessibles par l'IAO ont emmené avec eux la complexité en matière de contenu, de logiciel et de communications électroniques, laquelle n'est peut-être pas bien appréciée de tous. La conversion à l'IAO a entraîné une nouvelle dépendance par rapport à l'informatique. Cette dépendance est l'un des défis les plus importants auxquels Statistique Canada doit faire face, puisque la technologie évolue à un rythme effrénée.

4.1 Charge de travail des intervieweurs

La mise en commun d'une infrastructure nécessite le partage par différentes enquêtes de ressources limitées, comme des intervieweurs formés équipés d'ordinateurs portatifs. Par conséquent, toute augmentation du nombre d'enquêtes ou de la quantité des données recueillies dans une enquête doit être assumée conjointement par l'ensemble des autres enquêtes. Il faut souligner que, souvent, les mêmes intervieweurs travaillent pour un grand nombre d'enquêtes, de sorte qu'ils peuvent se retrouver avec une charge de travail considérable, situation exacerbée par la brièveté des périodes de collecte. Bien que le taux de réponse se soit rétabli depuis l'introduction de l'IAO, une charge de travail trop lourde peut altérer la qualité des données (moins de suivis et plus de non-réponses).

Compte tenu de la nature du SGC, il faut mettre en place une structure administrative à l'égard des communications, fondée sur les besoins de chaque enquête (selon les codes de réponses), pour permettre le cheminement des cas entre

3.2.2 Accès à des instruments de collecte plus perfectionnés

L'IAO a également donné accès à des instruments de collecte plus perfectionnés. Par exemple, dans le cadre de l'ELNEJ, on obtient une variété de renseignements sur une cohorte d'enfants âgés de 0 à 11 ans. Une section de l'interview consiste à évaluer le niveau de vocabulaire de l'enfant. L'un des instruments utilisés à cet égard est le test de vocabulaire par l'image de Peabody (PPVT). Toutefois, on utilise généralement le PPVT dans un environnement plus spécialisé, et les personnes qui administrent ce test doivent normalement suivre plusieurs jours d'une formation approfondie, le test nécessitant la présentation d'une série d'images parmi lesquelles l'enfant doit choisir celle qui correspond à un mot donné. Le niveau de départ du test dépend de l'âge de l'enfant. L'intervieweur pose des questions jusqu'à ce que l'enfant ait donné un certain nombre de mauvaises réponses. À ce moment, l'intervieweur doit retourner au niveau de départ et reposer les questions déjà posées, jusqu'à ce que l'enfant donne un nombre prédéterminé de mauvaises réponses. Pour administrer le test, il faut donc établir un seuil d'après certains critères, compter le nombre de mauvaises réponses, sauter des questions dans le cas où l'enfant donne un certain nombre de mauvaises réponses et mettre un terme au test. Cette marche à suivre aurait nécessité une formation très approfondie s'il avait fallu faire passer ce test sur papier. L'IAO a grandement facilité le procédé en permettant la préprogrammation des règles de validation. Les données de la première collecte permettent de penser que l'administration de ce type de test dans un environnement IAO offre des résultats de bonne qualité lorsqu'on les compare avec les normes externes.

3.2.3 Établissement de liens longitudinaux

Dans le cas des liens longitudinaux, il peut arriver que tous les membres d'un ménage initial fassent partie de l'échantillon longitudinal, à l'EDTR par exemple. Au cours des collectes suivantes, les personnes longitudinales sont interviewées, de même que toutes les personnes avec qui elles vivent. Si un ménage se sépare, on doit créer un nouveau ménage pour les personnes qui ont quitté le ménage d'origine. Grâce à l'adoption de l'IAO, il est devenu possible de créer des identificateurs de ménages propres aux nouveaux ménages mais reliés aux identificateurs originaux, et de retracer ainsi plus facilement la dynamique des changements dans la composition des ménages. Le traitement des doubles véritables qui résultent d'un changement dans la composition d'un ménage est un problème particulier qui a été grandement amélioré. Par exemple, un adolescent peut faire partie d'un ménage donné au moment de la première collecte, puis avoir laissé ses parents au moment de la deuxième interview, puis y être retourné quand arrive la troisième collecte. À la deuxième collecte, on indique que la personne fait partie d'un nouveau ménage et un nouvel identificateur y est associé. Lorsque l'on communique à nouveau avec les parents au moment de la troisième interview, l'adolescent qui est revenu pourrait passer pour un nouveau membre du ménage. Si l'intervieweur dispose de la liste des personnes qui ont déjà fait partie du ménage, la nécessité de réduire les doubles est grandement réduite. On a mis sur pied un procédé semblable dans le cas des emplois occupés par une personne, de sorte que la liste des employeurs précédents de celle-ci est utilisée pour une réconciliation longitudinal des emplois.

3.2.4 Dépistage des individus

Avec l'adoption de l'IAO, certaines fonctions ont pu être informatisées, notamment le dépistage. Brown et coll. (1997) en donnent des exemples précis. Comme on l'a noté plus haut relativement à l'établissement des liens longitudinaux, on peut inclure tous les individus «dépistes» dans un nouveau ménage en leur accolant un identificateur unique. Il y a moins de manipulation de papier, et il est maintenant possible d'obtenir davantage d'information en matière de gestion. Grâce à l'IAO, il a été possible de mettre en place une méthode de dépistage à deux niveaux. L'intervieweur essaie d'abord d'effectuer le dépistage. S'il ne réussit pas, toute l'information sur le cas est transférée à une unité de dépistage au bureau régional, où davantage de sources de dépistage sont disponibles. L'automatisation a éliminé de nombreuses manipulations et la transcription des données sur papier. Auparavant, lorsqu'un ménage se séparait, on devait créer sur papier une nouvelle feuille d'identification assortie d'un lien avec le ménage antérieur. Le nom des personnes qui avaient quitté le ménage était indiqué sur cette feuille. Si on ne trouvait pas la personne que l'on cherchait, il fallait transférer toutes les feuilles de toutes les personnes ayant vécu ensemble au cours de l'année précédente. Ces manipulations augmentaient considérablement le risque d'erreurs. Le transfert des cas entre les niveaux de dépistage se fait également plus rapidement. De plus, chaque recherche est enregistrée automatiquement avec son résultat. Même si la méthode était semblable à l'époque du crayon et du papier, il était rare que les renseignements soient enregistrés. Il était également difficile d'analyser l'information pour déterminer quelles seraient les meilleures sources pour retracer une personne.

Le dépistage est un facteur clé du maintien de la qualité des données. Grâce aux méthodes de dépistage actuelles, les cas devant faire l'objet d'une recherche peuvent demeurer sur le terrain un peu plus longtemps, même si la période de collecte demeure limitée. Il sera possible d'instaurer des méthodes plus efficaces si les efforts associés aux différentes enquêtes sont mis ensemble. On étudie actuellement comment atteindre une meilleure fonctionnalité, conjuguée à un dépistage centralisé. On pourrait ainsi combiner les efforts de dépistage des différentes enquêtes, et l'on pourrait aussi procéder à des saisies par lots afin de tenter de relier les cas nécessitant des recherches dans les bases de données.

particulier pour les longs questionnaires. Dans certains cas, l'information additionnelle pouvait même être imprimée sur un questionnaire séparé. Cette méthode posait d'autres problèmes logistiques pour l'intervieweur. L'utilisation d'information provenant d'interviews précédentes est connue sous le nom de rétroaction. Avec l'interview assistée par ordinateur, la rétroaction est rendue possible de deux manières: proactive et réactive. On trouvera un autre exposé sur ce sujet dans Brown et coll. (1997).

L'utilisation proactive de la rétroaction permet de réduire les erreurs de réponse en aidant le répondant à se situer. Par exemple, dans le cadre de l'EDTR, on recueille des renseignements détaillés sur un maximum de six emplois au cours de l'année précédente. Sans la rétroaction, le nom de l'employeur ou le titre du poste pourrait être écrit de façon légèrement différente et un emploi qui s'est poursuivi pendant deux ans pourrait être classé comme un changement. Au début, on a craint que les répondants perçoivent la rétroaction de façon négative, mais en fait, peu de commentaires négatifs ont été exprimés.

Le taux de confirmation est généralement élevé – plus de 90 % – pour les données qui sont présentées au répondant (voir Hale et Michaud 1995). L'étude de Hiemstra, Lavigne et Webber (1993) portant sur le marché du travail suggère que la rétroaction sert généralement à réduire les problèmes de concordance, mais que ceux-ci ne sont que partiellement résolus. Ainsi, dans le cadre de l'EDTR, on confirme l'occupation d'un emploi, la recherche d'un emploi, l'absence d'emploi au début de l'année civile précédente et pour une période d'un an pour laquelle le répondant doit faire appel à sa mémoire. Des micro- comparaisons avec une enquête transversale mensuelle menée au cours des cinq premiers mois de l'année ont permis d'observer que la rétroaction réduit considérablement les problèmes de concordance. Toutefois, la cohérence avec les données transversales diminue à mesure que les mois passent, ce qui laisse supposer que les erreurs de réponse, même si elles sont réduites par la rétroaction, continuent d'être un problème.

L'utilisation proactive de la rétroaction peut, cependant, créer une sous-estimation des mesures de changement. Pour cette raison, dans le cas d'information délicate ou pour des raisons de confidentialité, la technique est également utilisée de façon réactive. On peut utiliser la rétroaction réactive pour repérer des changements insolites, ou pour vérifier des incohérences dans les données. Par exemple, lors de l'interview de la première vague de l'EDTR, on demande au répondant d'indiquer ses périodes de chômage et, pour chaque période, s'il a reçu des prestations d'assurance-emploi. Au cours de l'interview de la deuxième vague, on demande des renseignements détaillés sur les différentes sources de revenu et les montants reçus, y compris les prestations d'assurance-emploi. Des comparaisons avec des sources externes ont permis d'établir qu'habituellement, les montants d'assurance-emploi rapportés dans une enquête représentent environ 80 % des prestations versées. Dans le cadre de l'EDTR, les renseignements précédents étaient conservés dans la mémoire de l'ordinateur. Si un montant n'était pas rapporté et qu'un indicateur signalait une incohérence avec la première interview, alors l'intervieweur posait une question additionnelle pour établir si le montant avait été omis. Une analyse de la première vague d'interviews de l'EDTR suggère que la rétroaction réactive a permis d'augmenter ce type de renseignements par une proportion de près de 30 %. Toutefois, 28 % des personnes qui avaient négligé de rapporter un montant de revenu ont confirmé qu'elles avaient bien reçu ce montant mais ont refusé d'en indiquer le montant. On pouvait donc confirmer la source du revenu, mais le montant devait être imputé et le problème n'était pas totalement résolu. Pour de plus amples renseignements sur ce sujet, consulter Dibbs, Hale, Loverock et Michaud (1995).

3.2 Un outil plus efficace

Grâce à un instrument de collecte aussi efficace que l'IAO, il est maintenant possible de recueillir des renseignements détaillés, de les limiter, d'y accéder et de les transférer, ce qui auparavant était très difficile, ou même impossible lorsque l'on utilisait le mode IPC.

3.2.1 Matrice des relations entre les différents membres d'un ménage

Les enquêtes-ménages créent différents niveaux d'analyse, tels que la famille économique et la famille de recensement, en utilisant les relations entre les différents membres du ménage et une personne appelée le «chef de famille». Cette méthode a ses limites, par exemple lorsqu'il s'agit d'identifier les enfants de familles mixtes ou de retracer une famille sur trois générations. Dans un contexte longitudinal, la définition de chef de famille peut varier avec le temps, et c'est pourquoi pour un certain nombre d'enquêtes on a utilisé une matrice des relations pour tous les membres du ménage. L'IAO peut limiter la collecte de données à la diagonale inférieure de la matrice. Si la composition d'un ménage n'a pas changé entre deux collectes de données, il n'est pas nécessaire d'établir à nouveau une matrice des relations. Les vérifications interactives (à propos de l'âge par exemple) servent à corriger toute relation saisie dans l'ordre inverse (par exemple une relation parent-enfant). On a dû procéder à un certain nombre d'essais pour élaborer un moyen efficace d'identifier les relations qui permettrait non seulement la collecte de renseignements mais leur correction facile. Grâce à la version améliorée de la méthode de collecte, moins de 1 % des relations ont besoin d'être corrigées après la collecte initiale (comparativement à un taux de 5,3 % d'incohérence avant les vérifications interactives sur la matrice des relations). Les méthodes de corrections des données dans un environnement d'IAO sont l'un des domaines où la recherche est encore nécessaire.

2.4 L'incidence de l'IAO sur la non-réponse

Y a-t-il lieu de croire que l'utilisation de l'IAO a eu un effet sur le taux de non-réponse? La réponse à cette question doit être affirmative, compte tenu des problèmes techniques survenus, principalement au début du processus de conversion. Cependant, si l'on fait abstraction de cet aspect, il ne semble pas que l'IAO ait un effet durable sur le taux de non-réponse. Dans le cas de l'EPA, le taux de non-réponse a fluctué à la suite de l'introduction de l'IAO, mais ces mouvements peuvent s'expliquer par un certain nombre d'autres facteurs (le remaniement de l'échantillon par exemple, qui est maintenant plus urbanisé ou l'embauche de nouveaux intervieweurs, *etc.*), puisque l'EPA a fait l'objet d'un remaniement majeur. Après juste un peu moins de deux ans, le taux de non-réponse est revenu à des niveaux semblables à ceux de la période du papier et crayon.

La conversion de l'EPA à la nouvelle méthode a pris cinq mois, au cours desquels on a pu comparer les taux de non-réponse des méthodes IPC et IAO. Ces comparaisons ont démontré que les taux de non-réponse de la méthode IAO (à l'exclusion des problèmes techniques) et ceux de l'IPC étaient du même ordre et suivaient les mêmes tendances (voir Simard et Dufour 1995). De plus, la répartition des principaux motifs de non-réponse, soit le refus de participer à l'enquête, l'absence temporaire du ménage, personne à la maison et autres raisons, était sensiblement la même avant et après l'adoption de la nouvelle méthode. On s'est inquiété que, dans le cas des interviews sur place, les répondants pourraient se montrer plus réticents à répondre eu égard à la présence de l'ordinateur, ce qui aurait fait croître le nombre de refus. Toutefois, on n'a pas détecté de variation quant à la composante refus de répondre.

Au début de 1995, la collecte des données des trois enquêtes longitudinales (EDTR, ELNEJ et ENSP) a été menée en même temps que celle de l'EPA. L'environnement de gestion de cas d'alors, conjugué à la mise en commun de l'infrastructure entre les enquêtes, a créé des pressions additionnelles sur les intervieweurs sur le terrain. De plus, les périodes de collecte des enquêtes étaient limitées parce qu'un nombre restreint d'applications pouvaient résider dans l'ordinateur en même temps. On a effectué une analyse pour déterminer si l'IAO provoquait un délai d'exécution provenant de la simultanéité ou de la succession rapide des enquêtes sur le terrain. Dans le cas de la collecte trimestrielle de l'ENSP, les intervieweurs faisaient une relance auprès des non-répondants des collectes antérieures. On a procédé à une analyse de cette opération pour évaluer le taux de conversion possible. Les résultats ont montré que, lorsque qu'il y avait moins d'enquêtes IAO sur le terrain en même temps, une première vague de relance des non-répondants augmentait le taux de réponse, mais que reproduire l'opération une deuxième ou une troisième fois n'apportait que peu de gains additionnels (augmentation de 5,76 % du premier au deuxième trimestre, de 0,97 % du deuxième au troisième et de 0,91 % du troisième au quatrième). Toutefois, une dernière relance fut effectuée en juin 1995, alors qu'il n'y avait pas presque d'autres enquêtes en cours. L'opération a permis de hausser le taux de réponse d'environ 5 %, ce qui était plus élevé que prévu. On en a conclu que l'IAO devait s'accompagner d'une plus grande souplesse relativement à la longueur de la période de collecte de données et qu'il fallait que plusieurs applications puissent résider dans l'ordinateur en même temps, de sorte que l'on puisse conserver les taux de réponse du temps de la méthode du papier et du crayon.

3. DE NOUVELLES POSSIBILITÉS POUR LES ENQUÊTES-MÉNAGES

L'adoption de l'interview assistée par ordinateur a ouvert de nouvelles possibilités en ce qui concerne les enquêtes-ménages. Ces nouvelles possibilités, qui étaient ou bien inexistantes ou difficiles à réaliser avec la méthode du papier et du crayon, permettent de réduire les erreurs non dues à l'échantillonnage, de recueillir des renseignements plus spécialisés, de faciliter la reconstruction des entités familiales et de joindre les éléments des unités familiales qui se sont séparées ou fusionnées. En fait, cette méthode de collecte est mieux adaptée aux besoins changeants de la société d'aujourd'hui.

3.1 Interviews dépendantes

L'introduction de la nouvelle technologie a permis de résoudre des problèmes qui s'étaient avérés insolubles lorsque les enquêtes-ménages étaient effectuées au moyen de la méthode du papier et crayon. Notamment, l'IAO a permis d'accroître la quantité d'information fournie par l'intervieweur à un répondant joint pour la seconde fois et i) de réduire les erreurs de réponse (erreur de codage, de saisie ou de mémoire), et particulièrement les problèmes de concordance et de télescopage et ii) d'alléger la tâche du répondant en confirmant les renseignements plutôt qu'en les demandant de nouveau (ou en n'en demandant qu'une partie).

Les problèmes de concordance ont été décrits pour les enquêtes longitudinales par Murray, Michaud, Egan et Lemaître (1990), qui explique qu'ils se produisent lorsque l'on essaie de réconcilier les données de périodes de collecte successives. Si l'on n'avait pas tenté de faire des réconciliations entre les collectes de données, on aurait observé généralement des variations artificiellement importantes entre les estimations provenant de deux périodes consécutives. Ce problème s'explique généralement du fait que les répondants ont de la difficulté à indiquer la date exacte d'un changement. En ce qui concerne le télescopage, il provient d'un tendance à inclure certains événements s'étant produits à l'extérieur de la période de référence.

Avec la méthode papier et crayon, les intervieweurs ne pouvaient disposer que d'une quantité limitée d'information. Les questionnaires ne pouvaient que contenir de l'information de base, puisqu'il y avait des limites à la quantité de renseignements pouvant être préimprimés, en

femmes. L'interview téléphonique assistée par ordinateur continue de faire partie intégrante du système de collecte des données auprès des ménages à Statistique Canada et de servir de complément à l'infrastructure de l'interview assistée par ordinateur.

2.2 Essais technologiques

Une nouvelle vague de tests a pris son essor au début des années 1990, dans le cadre du remaniement décennale de l'EPA (Singh, Gambino et Laniel 1993; Drew, Gambino, Akyeampong et Williams 1991). Grâce au lancement de trois enquêtes longitudinales à grande échelle qui permettait une mise en commun des coûts, Statistique Canada a pu engager les fonds pour la mise en place d'une infrastructure d'IAO. En 1991, on a donc procédé à un deuxième essai sur l'EPA et l'EDTR pour évaluer la faisabilité d'utiliser les nouvelles technologies (voir Williams et Spaull 1992). On a fait l'essai des ordinateurs portatifs, qui fonctionnent avec un stylet plutôt qu'un clavier pour la saisie des données. Les résultats ont montré que la technologie était prometteuse mais qu'il y avait matière à amélioration avant qu'elle puisse répondre aux exigences se rapportant à la conduite des enquêtes-ménages à Statistique Canada.

L'année suivante, de juillet 1992 à janvier 1993, on a effectué un troisième et un quatrième essais, mais cette fois au moyen d'ordinateurs portatifs conventionnels. Les résultats pour l'EPA sont présentés dans Kaushal et Laniel (1995), tandis que les résultats pour l'EDTR sont rapportés dans Michaud, Le Petit et Lavigne (1993) et Michaud, Lavigne et Pottle (1993). Dans le cas de l'EPA, le troisième essai avait pour principal objectif d'établir si une conversion à la nouvelle technologie aurait pour effet de perturber la série de données de l'EPA. L'objectif secondaire était de déterminer si la nouvelle technologie influencerait la qualité des données et les frais d'interview. Il s'agissait également de procéder au développement opérationnel et à l'évaluation de l'IAO. Pour ce qui concerne les enquêtes longitudinales, la principale préoccupation était la longueur et la complexité des questionnaires et l'ajout de nouvelles fonctions comme le dépistage. Par conséquent, le principal critère d'évaluation de l'application était la faisabilité de développer diverses fonctions. Les résultats ont montré que l'IAO n'avait pas d'influence importante pour l'EPA que ce soit sur la diffusion de la série de données, sur les principaux indicateurs de qualité ou sur les coûts d'interview. Après des comparaisons générales avec des sources externes et une analyse des variables manquantes, on a adopté la nouvelle technologie.

2.3 Nouvelle dimension de la non-réponse

L'adoption de l'IAO a entraîné l'apparition impromptue d'une nouvelle dimension de non-réponses causées par des «problèmes techniques». Ces non-réponses provenaient de cas perdus ou non reçus avant la fin de la période de collecte. Ce type de non-réponse existait avec la méthode IPC sous la forme de problèmes postaux occasionnels. Conceptuellement, ces situations ne se rapportent pas à de véritables refus de répondre; toutefois l'information n'est pas disponible à temps pour faire partie des estimations.

Ces problèmes techniques peuvent prendre trois formes différentes: i) problèmes de transmission, ii) problèmes matériels et iii) problèmes inévitables. Les problèmes de transmission sont les plus courants. Ils se produisent, par exemple, lorsque les lignes téléphoniques sont en panne, lorsqu'il y a une difficulté empêchant le téléchargement automatique des données, lorsque l'on tente de télécharger les données au moment où l'ordinateur central fait l'objet de travaux de maintenance, ou simplement parce qu'il y a un mauvais fonctionnement du système IAO. Le second type de problème, qui est moins courant, arrive lorsqu'un disque dur ou un lecteur de bande magnétique tombe en panne, qu'il y a insuffisance de mémoire ou qu'il y a un problème de matériel informatique au bureau régional. Enfin, les problèmes inévitables, qui sont encore plus rares, sont des problèmes particuliers implicitement causés par l'une des situations ci-dessus, par exemple lorsque seulement l'une des deux composantes des réponses d'un sondé est transmise ou si les paramètres d'initialisation nécessaires au bon fonctionnement des programmes font défaut.

Le nombre de non-réponses attribuables à des problèmes techniques a diminué au cours des premiers mois. On a analysé très soigneusement cette composante de la non-réponse pour expliquer la tendance à la hausse à cet égard et pour évaluer la performance de la méthode IAO (voir Simard, Dufour et Mayda 1995, Dufour, Simard et Mayda 1995). Au début de la conversion des enquêtes- ménages à l'IAO, les problèmes techniques représentaient en moyenne 15 % du nombre total de non-réponses et pouvaient expliquer jusqu'à 25 % de celles-ci. Ce n'est qu'environ au bout d'une année entière que l'on a pu observer une réduction importante de cette composante de la non-réponse. Aujourd'hui en 1997, les non-réponses attribuables à des problèmes techniques sont à peu près inexistantes.

Au cours de la première année, le gros des problèmes était causé par un conflit de gestion de mémoire dans l'ordinateur portatif entre deux logiciels servant à la gestion des cas. On élimina le conflit en réécrivant une partie du logiciel, ce qui rendit le système plus efficace. Les éléments les plus subtils de cette période de transition étaient la communication et l'expérience. On a élaboré une stratégie de communication pour permettre aux différents intervenants (en particulier le personnel technique et les intervieweurs) de mieux comprendre le rôle de chacun, de diffuser l'information plus rapidement et d'informer adéquatement toutes les personnes concernées. Lorsque l'IAO a été introduite initialement, certains problèmes prenaient plus d'un jour avant d'être résolus par le personnel de soutien technique. Des procédures visant à accélérer le dépannage ont été élaborées, et un service de soutien de 24 heures a été mis en place au siège social à Ottawa. Dans le cas d'un changement aussi important, une période d'apprentissage et d'ajustement est nécessaire, et, à Statistique Canada, on n'a pas fait exception à cette règle.

amples détails sur la structure et la mise en oeuvre de cette méthode de collecte informatisée dans le cadre des enquêtes longitudinales, voir Brown, Hale et Michaud 1997. Aujourd'hui, la plupart des données des enquêtes-ménages de Statistique Canada sont collectées par technique informatisée et partagent une infrastructure commune.

Cet article porte surtout sur les aspects méthodologiques de l'interview assistée par ordinateur dans un milieu décentralisé telle qu'elle a été appliquée aux enquêtes-ménages. On présente une vue d'ensemble du processus de mise en oeuvre à Statistique Canada dans son ensemble, une brève présentation des défis associés à cette nouvelle méthode de collecte et une bibliographie pour permettre au lecteur d'en apprendre davantage sur certains sujets précis. Malgré les difficultés de croissance, Statistique Canada continue d'expérimenter et de mettre en oeuvre cette nouvelle technologie dans le cadre de différentes enquêtes afin d'améliorer leur rapport coût-efficacité, la qualité des données et le processus de suivi de ces enquêtes.

Cet article comprend cinq sections. Dans la section suivante, on présente divers aspects de la mise en oeuvre de l'interview assistée par ordinateur dans le cadre de différentes enquêtes. La section 3 présente les nouvelles possibilités offertes par l'IAO. Dans la section 4, on passe en revue les enjeux actuels et les nouveaux problèmes auxquels les enquêtes doivent faire face suite à l'application de cette méthode de collecte informatisée, de même que les changements qui en découlent. La dernière section évoque l'avenir de l'IAO pour les enquêtes- ménages à Statistique Canada.

2. LES PREMIÈRES ANNÉES DE MISE EN OEUVRE

L'adoption d'une méthode de collecte informatisée pour les enquêtes-ménages offrait plusieurs avantages prometteurs: i) une réduction des coûts d'enquête, ii) une meilleure qualité des données, iii) la possibilité d'utiliser des questionnaires plus complexes, iv) des données disponibles plus rapidement, v) un outil de dépistage, vi) la possibilité de réaliser des interviews dépendantes et vii) une méthode de collecte généralisée pour toutes les enquêtes-ménages de Statistique Canada. Toutefois, ces avantages ne se sont pas concrétisés du jour au lendemain ou sans effort. Il a fallu, au cours des étapes d'introduction et de stabilisation, procéder à des évaluations et des rajustements constants.

Bien qu'un certain nombre d'essais aient été effectués avant la mise en oeuvre de l'IAO, l'adoption de cette méthode a entraîné des problèmes imprévus, même si, avec le temps, ils sont devenus moins nombreux et plus faciles à résoudre. De plus, au cours de cette période, la série d'indicateurs de la qualité analysés soigneusement par différents groupes d'experts de Statistique Canada a été quelque peu perturbée. Les avantages anticipés ont pris environ un an avant de se réaliser. La présente section décrit les principaux points du passage entre la méthode traditionnelle «sur papier» à la méthode d'interview assistée par ordinateur, laquelle permet d'intégrer la collecte et la saisie des données.

2.1 L'interview téléphonique assistée par ordinateur en environnement centralisée

La méthode traditionnelle d'interview consistait à utiliser un questionnaire sur papier que l'intervieweur remplissait avec un crayon afin de faciliter les corrections. On fait souvent référence à cette méthode sous l'appellation «interview papier et crayon (IPC)». Avec cette méthode traditionnelle, l'intervieweur vérifiait le questionnaire pour s'assurer que les renseignements consignés étaient exacts et complets. Les abréviations utilisées pour réduire la durée de l'interview étaient retranscrites au long après l'interview avant que le questionnaire soit transmis pour la saisie des données. La première étape vers l'informatisation a été franchie avec l'adoption de l'«interview téléphonique assistée par ordinateur» (ITAO). On utilisait cette méthode de collecte de données pour les enquêtes menées par téléphone à partir d'un emplacement unique. L'ITAO a été la première expérience d'intégration de la collecte et de la saisie d'information dans le cadre des enquêtes-ménages. Compte tenu de la technologie de l'époque, il fallait utiliser des ordinateurs de taille relativement considérable pour traiter la complexité associée à l'interview assistée par ordinateur. Il n'était donc possible de remplacer l'IPC par l'ITAO que dans le cadre d'enquêtes téléphoniques centralisées. Dans les années 1990, l'avènement d'ordinateurs portatifs plus puissants a permis à l'IAO en milieu décentralisé de remplacer l'IPC. On a en effet maintenant recours à une méthode de collecte décentralisée pour la plupart des enquêtes-ménages. De plus, cette collecte décentralisée requiert souvent que l'interview puisse se faire par téléphone ou en personne. Quoi qu'il en soit, la plus grande part du savoir-faire et de l'expérience acquis avec l'interview téléphonique assistée par ordinateur a pu être appliquée à l'interview assistée par ordinateur dans un environnement décentralisé.

Depuis les années 1980, c'était l'Enquête sur la population active (EPA) qui servait pour la recherche et les essais technologique du monde ITAO. Le premier essai a été effectué en 1987 sous la forme d'une étude contrôlée qui comparait l'ITAO dans un environnement centralisé avec l'IPC. Il s'agissait d'un projet de recherche mené conjointement par Statistique Canada et le Bureau of the Census des États-Unis (voir Catlin et Ingram 1988). L'étude a mis en relief les écarts qui existaient entre les méthodes du point de vue de la qualité des données, ces différences favorisant l'IAO (réduction du taux de rejet lors des vérifications, réduction des erreurs d'aiguillage sur les questionnaires et diminution du sous-dénombrement à l'égard de l'EPA).

Bien que l'ITAO n'ait jamais été appliquée à l'EPA, l'expérience a servi à mettre au point une fonction ITAO de composition aléatoire (CA) pour les enquêtes-ménages. Avec l'évolution technologique, l'ITAO a servi à des enquêtes CA plus complexes comme l'Enquête sociale générale (ESG) et l'Enquête sur la violence envers les

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 159-169
Statistique Canada

Les interviews assistées par ordinateur dans un environnement décentralisé: Le cas des enquêtes-ménages à Statistique Canada

J. DUFOUR, R. KAUSHAL et S. MICHAUD[1]

RÉSUMÉ

En 1993, Statistique Canada introduisait l'Interview assistée par ordinateur (IAO) pour certaines enquêtes-ménages menées dans un environnement décentralisé. Cette technologie a été utilisée avec succès pendant quelques années et la plupart des enquêtes-ménages sont maintenant converties à cette méthode de collecte. Le présent document fait un résumé de l'expérience acquise et des leçons apprises depuis le début de la recherche sur le sujet. Il décrit certains des essais qui ont mené à l'adoption de cette technologie et quelques-unes des nouvelles possibilités qui sont nées de sa mise en oeuvre. Il présente aussi un certain nombre d'enjeux qui se sont posés lors de l'adoption de l'IAO (certains existant encore aujourd'hui) et se termine sur un bref survol de ce que nous réserve l'avenir.

MOTS CLÉS: Enquêtes-ménages; collecte de données; interviews assistées par ordinateur; environnement décentralisé.

1. INTRODUCTION

Les premiers systèmes d'interview assistée par ordinateur (IAO) ont été mis au point au début des années 1970 (voir Nicholls et Groves 1986). Ces systèmes ont surtout été élaborés par des organisations faisant des études de marché aux États-Unis et, un peu plus tard et de façon indépendante, par des centres de recherche universitaires bien connus. Vers la fin de 1a décennie 1970 et le début des années 1980, les systèmes d'interview assistée par ordinateur se sont considérablement perfectionnés, et leur usage s'est largement répandu. Ainsi, vers la fin des années 1980, un nombre des universités et centres d'enquête américains possédaient un système de collecte informatisée (voir Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz et Trewin 1997). Clark, Martin et Bates (1997) font un survol de l'élaboration et de la mise en oeuvre de ces systèmes dans quatre grandes organisations statistiques gouvernementales.

En 1987, Statistique Canada faisait ses premiers essais en matière d'interview assistée par ordinateur en l'appliquant aux enquêtes-ménages. Les essais avaient alors lieu dans un «milieu centralisé de collecte des données par téléphone». Cette série d'essais a été prolongée jusqu'au début des années 1990, dans un effort pour adapter cette technologie aux méthodes plus générales de collecte de données.

La plupart des enquêtes-ménages effectuées à Statistique Canada partagent la même base de sondage et le même environnement de collecte de données. Le principal utilisateur de cette base est l'Enquête sur la population active (EPA) mensuelle. La collecte des données est décentralisée, la première interview ayant lieu sur place au logement du ménage choisi et les cinq interviews suivantes étant menées par téléphone à partir de la résidence de l'intervieweur. Pour ce faire, près d'un millier d'intervieweurs ont été équipés d'un ordinateur portatif. Les intervieweurs sont rattachés à l'un des cinq bureaux régionaux couvrant le Canada. Statistique Canada adopte une stratégie similaire pour un certain nombre d'enquêtes-ménages, en procédant à un sous-échantillonnage de l'échantillon de l'Enquête sur la population active, en administrant une série de questions supplémentaires après l'interview de l'EPA proprement dite ou en communiquant avec des personnes qui ont déjà participé à l'enquête. Par conséquent, l'EPA partage avec les autres enquêtes non seulement son échantillon mais aussi son infrastructure de collecte des données. Tous les intervieweurs sont tenus de travailler pour le compte de l'EPA pendant une semaine précise de chaque mois, tandis que, le reste du temps, ils se consacrent à d'autres enquêtes, ayant été équipés et formés en ce sens. Pour de plus amples renseignements sur la méthodologie de l'Enquête sur la population active, voir Statistique Canada (1998).

Dans les années 1990, on a étendu l'essai de la méthode de collecte assistée par ordinateur non seulement à l'EPA mais également à d'autres enquêtes, qui partageaient une infrastructure commune mais avaient des besoins très différents. Les résultats de ces différents essais ont mené à la mise en oeuvre, en novembre 1993, de l'interview assistée par ordinateur dans le cadre de l'EPA (Dufour, Kaushal, Clark et Bench 1995), tandis que les enquêtes mensuelles supplémentaires de l'EPA étaient graduellement modifiées par la suite. En janvier 1994, une nouvelle enquête longitudinale, l'Enquête sur la dynamique du travail et du revenu (EDTR) était lancée, laquelle recourait à l'interview assistée par ordinateur (voir Lavigne et Michaud 1995). Depuis lors, l'Enquête nationale sur la santé de la population (ENSP) et l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), lancées en août et novembre 1994 respectivement, adoptaient également cette méthode de collecte (voir Tambay et Catlin 1995, Brodeur, Montigny et Bérard 1995). Pour de plus

[1] J. Dufour et R. Kaushal, Division des méthodes d'enquêtes des ménages; S. Michaud, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, (Ontario), K1A 0T6.

une étape importante du travail sur le terrain. L'exécution minutieuse de ces tâches permet de retracer les unités échantillonnées aux fins des suivis qui seront une activité de mesure indispensable pour évaluer le projet IFPS.

Le fait qu'une enquête aussi complexe que PERFORM, exécutée à une échelle qui permet de saisir tant les niveaux que les variations de la prestation de services de santé et d'utilisation des services par les clients dans une région aussi peuplée que l'Uttar Pradesh, produise des données qui satisfont la plupart des normes de précision témoigne, sans conteste, d'un grand accomplissement sur le terrain, ainsi que d'une innovation importante en matière de plan d'échantillonnage.

REMERCIEMENTS

La présente étude a été financée en partie par The EVALUATION Project, USAID Contract #DPE-3060-C-00-1054-00. Les opinions exprimées ici n'engagent que les auteurs et ne représentent pas celles de l'organisme parrain. Les auteurs expriment leur gratitude à Daniel Horowitz et à T.K. Roy pour l'aide qu'ils leur ont apportée antérieurement pour établir le plan d'échantillonnage. Ils remercient aussi Lynn Moody Igoe, du Carolina Population Center, d'avoir révisé l'article. Enfin, ils remercient les examinateurs anonymes de leurs suggestions et de leurs commentaires précieux.

BIBLIOGRAPHIE

ADAY, L.A. (1991). *Designing and Conducting Health Surveys: A Comprehensive*. San Francisco: Jossey-Bass Publishers.

BOYD, L.H., Jr., et IVERSION, G.R. (1979). *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth.

MACRO INTERNATIONAL, INC. (1996). *Demographic and Health Surveys Newsletter*, 8, 1-12.

MILLER, R.A., NDHIOVU, L., GACHARA, M.M., et FISHER, A.A. (1991). The situation analysis study of the family planning program in Kenya. *Studies in Family Planning*, 22, 131-143.

NARAYANA, G., CROSS, H.E., et BROWN, J.W. (1994). Family planning programs in Uttar Pradesh issues for strategy development: tables. Centre for Population and Development Studies, Hyderabad, India.

ROSS, J.A., et McNAMARA, R. (Éds.) (1983). *Survey Analysis for the Guidance of Family Planning Programs*. Liege, Belgium: Ordina Editions.

Tableau 6
Nombre échantillonné observé et prévu de CCS/CPS[a] et de sous-centres dans les villages ruraux (îlots urbains), selon le district, Uttar Pradesh (Inde), 1995

District	CCS/CPS		Sous-Centre		Organisme chargé du travail sur le terrain
	Réel	Estimé	Réel	Estimé	
Aligarh	6	5	10	17	II
Azamgarh	3	5	24	15	III
Almora	5	2	14	9	I
Allahabad	19	4	17	18	III
Ballia	9	7	34	27	III
Banda	8	9	19	27	III
Bareilly	5	3	10	16	II
Dehradun	5	7	10	21	I
Etawah	8	7	17	20	II
Fatehpur	9	7	22	25	IV
Firozabad	6	6	28	30	II
Gonda	8	5	15	18	IV
Gorakhpur	5	4	16	20	IV
Jhansi	7	6	16	24	II
Kanpur	2	2	6	8	II
Maharajgang	4	4	9	13	IV
Meerut	12	8	12	34	II
Mirzapur	7	7	22	22	III
Moradabad	5	5	9	19	I
Nainital	6	4	19	19	I
Rampur	2	5	14	16	I
Saharanpur	6	6	25	21	I
Shahjahanpur	5	3	14	15	II
Sultanpur	16	6	21	15	IV
Tehri	1	3	3	10	I
Unnao	3	6	17	17	IV
Sitapur	10	6	9	24	IV
Varanasi	6	5	18	18	III
Total	186	147	450	538	
Total[b]	151	137			

[a] Inclut les centres primaires de santé supplémentaires
[b] N'inclut pas les districts d'Allahabad et de Sultanpur.

Parallèlement, plusieurs enseignements se dégagent de notre application du plan d'enquête proposé. Premièrement, il est manifeste qu'il faut surveiller étroitement le travail sur le terrain et intensifier la saisie de données sur place, afin d'empêcher le phénomène apparent consistant à«pousser» des femmes admissibles hors des groupes d'âge les plus avancés. Ce phénomène est difficile à déceler par vérification ponctuelle des questionnaires individuels, mais peut être dépisté grâce aux totalisations agrégées produites, disons, hebdomadairement d'après les questionnaires remplis. Deuxièmement, le surdénombrement des CCS/CPS dans deux districts, où le travail sur le terrain a été effectué par deux organismes distincts, donne à penser que les villages de la strate I ont été sélectionnés de façon disproportionnée ou que certains CCS/CPS déclarés comme étant dans les limites de l'USÉ ne l'étaient pas en réalité. La première situation peut avoir eu lieu à cause d'une erreur d'échantillonnage, puisque chaque organisme chargé du travail sur le terrain a reçu une liste des USÉ échantillonnées. Troisièmement, le listage et le relevé cartographique des établissements, des prestateurs privés de services de santé et des ménages à l'échelle des USÉ est

ainsi le nombre prévu de CCS/CPS et de SC dans chaque district. Puis, nous avons comparé les résultats obtenus aux chiffres recueillis pour ce type d'établissements au moment du travail sur le terrain auprès des informateurs communautaires auxquels on a demandé d'indiquer s'il existait un CCS/CPS et (ou) des SC dans l'USÉ. La comparaison est présentée au tableau 6, qui montre aussi le code de l'organisme chargé du travail sur le terrain (I à IV) permettant de repérer toute erreur systématique éventuelle d'enquête. En appliquant cette méthode, on surestime le nombre de sous-centres de 19,6 % et on sous-estime le nombre de CCS/CPS de 26,5 %. Si on élimine les deux districts comptant un grand nombre d'USÉ dans la strate I (Allahabad et Sultanpur), la surestimation du nombre de CCS/CPS n'est plus que de 10,2 %. La totalisation de l'erreur d'estimation selon l'organisme du travail sur le terrain n'indique aucun biais.

Les résultats des deux méthodes de pondération donnent à penser que l'USÉ donne une mesure de population appropriée pour la sélection des sous-centres, puisque la taille moyenne de sa population s'approche de l'effectif des secteurs desservis par les SC, soit 5 500. Une mesure plus grande de population aurait sans doute donné de meilleurs résultats dans le cas de la sélection des CCS/CPS, puisque le secteur desservi par ces établissements couvre ceux desservis par cinq à six sous-centres. Comme on s'appuie sur la taille de l'USÉ pour calculer le coefficient de pondération du CCS/CPS, si l'USÉ est petite, le biais qui entache les dénombrements estimés peut être important. Un plan de sondage qu'il conviendrait d'étudier dans l'avenir consiste à sélectionner une grappe d'USÉ contiguë à l'USÉ sélectionnée pour obtenir une mesure d'effectif comparable à la population du secteur desservi par les CCS/CPS. Alors, la probabilité que pareil établissement se situe dans les limites de la grappe d'USÉ sera plus forte et le poids, calculé d'après le total de la population de la grappe d'USÉ sera plus fiable. Autrement dit, le fait de ne pas savoir combien d'USÉ sont desservies par un CCS/CPS limite la précision de l'estimation.

4. DISCUSSION

Le plan d'échantillonnage en grappes pour la production d'échantillons indépendants d'établissements et de ménages que l'on peut analyser individuellement ou collectivement mérite d'être considéré davantage pour la collecte des données nécessaires à l'étude et à l'évaluation des programmes de santé dans les pays en voie de développement. Si on fait preuve de minutie pour établir le plan d'enquête et pour exécuter ce dernier sur le terrain, on obtient des estimations par sondage de grande qualité et de précision acceptable, comme l'indiquent nos résultats. Les totaux pondérés, plutôt que les totaux d'échantillon représentent eux-mêmes des chiffres utiles pour les planificateurs de programme qui doivent décider des flux de personnel, de matériel et de fonds vers les divers établissements et prestateurs de soins locaux et entre ces derniers. De surcroît, le couplage d'un établissement à des enregistrements individuels offre d'importantes possibilités analytiques, comme l'évaluation de l'importance relative de facteurs liés aux antécédents professionnels du personnel et à la fourniture de services sur les résultats particuliers étudiés en matière de santé (p. ex., Boyd et Iversion 1979).

Tableau 5

Nombre total réel et estimé de centres communautaires de santé, de centres primaires de santé[a] et de sous-centres, selon le district, Uttar Pradesh (Inde), 1995

District	CCS/CPS		Sous-centre	
	Réel	Estimé	Réel	Estimé
Aligarh	77	69	399	369
Azamgarh	103	69	475	949
Almora	44	104	254	468
Allahabad	112	981	594	677
Ballia	73	93	357	485
Banda	89	101	322	302
Bareilly	71	42	355	162
Dehradun	24	41	139	60
Etawah	69	84	323	364
Fatehpur	57	73	309	327
Firozabad	33	34	234	236
Gonda	107	183	528	461
Gorakhpur	59	84	470	460
Jhansi	51	77	251	157
Kanpur Nagar	12	13	81	74
Maharajgang	30	39	195	180
Meerut	76	187	410	119
Mirzapur	64	69	309	302
Moradabad	92	81	485	248
Nainital	53	79	287	344
Rampur	37	19	170	139
Saharanpur	60	49	293	388
Shahjahanpur	52	59	301	298
Sultanpur	70	487	394	649
Tehri Garhwal	31	5	159	63
Unnao	63	162	344	106
Sitapur	87	44	437	450
Varanasi	122	144	616	658
Total	1 818	3 472(±21)	9 491	9 495(±15)
Total[b]	1 636	2 004(±13)		

[a] Inclut les centres primaires de santé supplémentaires
[b] N'inclut pas les districts d'Allahabad et de Sultanpur
Source des chiffres réels de 1995: gouvernement de l'Uttar Pradesh, ministère de la Santé et du Bien être familial.

112 568 villages, ce qui donne à penser qu'il existait pratiquement une accoucheuse traditionnelle par village et un travailleur anganwandi pour 4,5 villages, en moyenne. Ces ratios semblent raisonnables compte tenu de ce que l'on sait de l'accès à ce genre de soin. Les chiffres sont fort comparables et prouvent qu'il est utile de se servir d'un plan d'échantillonnage en grappes enchaînées.

3.3 Méthodes d'estimation

Les nombres estimés de CCS/CPS et de SC présentés au tableau 4 se fondent sur l'hypothèse selon laquelle pareils établissements desservent une population de taille constante, c.-à-d. 30 000 personnes et 5 500 personnes, respectivement, chiffres qui sont ceux utilisés par l'administration publique pour planifier la fourniture de services de santé. La précision des estimations serait meilleure si on connaissait la taille réelle de la population des secteurs desservis. Faute de ces renseignements, nous avons choisi une estimation constante de population pour ces deux types d'établissements.

Nous avons examiné d'autres méthodes d'estimation avant de choisir celle susmentionnée. La première est illustrée au tableau 5 où sont présentés les nombres réels et pondérés de CCS/CPS et de SC dans chacun des 28 districts observés. Ces chiffres se fondent sur la pondération des établissements sélectionnés selon la taille de l'USÉ uniquement, sans correction pour tenir compte de la multiplicité. L'échantillon PERFORM compte en tout 633 CCS/CPS, soit 34.8 % du total (1 818), et 1 267 SC; soit 13.3 % du total (9 491), sélectionnés dans les 28 districts. Si on compare ces chiffres au nombre réel de CCS/CPS et de SC relevés en 1995 par le ministère de la Santé et du Bien-être familial de l'Uttar Pradesh, on constate que la méthode de pondération susmentionnée aboutit à une surestimation importante du nombre de CCS/CPS (3 472 comparativement à 1 818), mais produit un nombre pratiquement identique de SC (9 495 comparativement à 9 491). L'utilisation des villages et des îlots urbains comme USÉ est raisonnable, puisqu'il s'agit des unités (et des chiffres de population) que l'administration publique utilise pour déterminer l'emplacement des sous-centres.

Cependant, ces unités ne représentent pas une base de stratification appropriée pour les grands établissements de santé. Il y a perte de précision, car, comme nous prenons pour poids l'inverse de la population de l'USÉ, ce poids est gonflé de façon disproportionnée quand on sélectionne des CCS/CPS dans des très petites USÉ. Il y a alors surdénombrement de ce type d'établissement, situation qui est particulièrement problématique dans deux districts – Allahabad et Sultanpur. Si on supprime ces deux districts, la surestimation est de 22.5 % (± 0,8) au lieu de 91 %. (Dans la situation inverse, comme c'est le cas pour le district de Bareilly, on aboutit à une sous-estimation des CCS/CPS. En raison de l'échantillonnage avec probabilité proportionnelle à la taille (PPT), les grands villages de la strate IV ont un faible poids et, en fait, la plupart des PFSF de ce district ont été sélectionnés pour des USÉ de cette taille.)

Une deuxième méthode d'estimation que nous avons utilisée consiste à calculer le nombre prévu de CCS/CPS et de SC en sachant a priori que les établissements de ce genre sont situés dans une USÉ dont la taille minimale est de 30 000 ou 5 500, respectivement. Grâce aux données du Recensement de 1991 sur la population des USÉ, nous avons reconstruit la courbe de répartition de la population de chaque district selon la taille de la strate et divisé chaque strate par la taille du secteur desservi par le CCS/CPS ou le SC (30 000 ou 5 500, respectivement). Nous avons obtenu

Tableau 4
Nombre total de points publics et privés de fourniture de services, selon le type, Uttar Pradesh (Inde), 1995

Points de fourniture de services fixes	Nombre	Prestateurs individuels de services	Nombre
Total	31 400	Total	1 099 825
Hôpitaux		Médecins particuliers	
Gouvernementaux- allopathie	968	Résidents-allopathie	32 182
Gouvernementaux-MIC	688	Agréés-allopathie	9 011
Municipaux-allopathie	57	Résidents (non qualifiés)	62 880
Municipaux-MIC	23	Résident-MIC	42 343
Privés	5 212	Agréés-MIC	9 138
Privés bénévoles	130	Travailleurs Anganwadi	25 994
Privés-MIC	35	Travailleurs de la santé des villages	65 532
Industriels	61	Accoucheuses traditionnelles	110 546
Écoles de médecine	9	Magasins de produits et services médicaux	40 979
CCS/CPS/CPS supplémentaires	3 948	Magasins de marchandises diverses	133 517
Sous-centres	20 151	Magasins Kirana	376 679
Autre	137	Bureaux de prêteurs sur gage	136 353
		Détenteurs de dépôts	5 818
		Autre	48 855

dans les deux groupes d'âge (de 0 à 14 ans et 65 ans et plus) est bonne, ainsi que celle des pourcentages de ménages appartenant aux castes désignées. La proportion des ménages appartenant aux tribus désignées est égale à 3,1, valeur supérieure à celle 1,1 observée dans le cas de la NFHS. Ces résultats pourraient refléter une croissance réelle du nombre de ces ménages, accompagnée d'une augmentation de l'immigration des membres des tribus désignées dans les grandes villes. La proportion de personnes sachant lire et écrire a augmenté légèrement depuis l'exécution de la NFHS, mais dans l'ensemble, les résultats sont comparables. L'indice synthétique de fécondité et le niveau d'utilisation des contraceptifs modernes sont également similaires et les directions de leur variation durant l'intervalle entre les deux enquêtes effectuées en Uttar Pradesh concordent. Les résultats du tableau 2 donnent à penser que le plan d'échantillonnage de l'enquête PERFORM, fondée sur un échantillonnage en grappes à plusieurs degrés utilisé ordinairement pour les enquêtes démographiques, a été exécuté comme il convient pour produire des résultats au niveau de l'État comparables à ceux du recensement et de la NFHS effectués antérieurement. Le tableau renseigne aussi sur l'érreur-type et sur l'effet du plan d'échantillonnage sur les estimations.

Au tableau 3, nous comparons la répartition de la population de l'Uttar Pradesh selon l'âge et le sexe établie d'après la NFHS et d'après l'enquête PERFORM, ainsi que d'après le Sample Registration System (système d'enregistrement des échantillons) tenu par le bureau général de l'état civil. Nous donnons aussi les rapports de masculinité calculés d'après les résultats des deux enquêtes. De nouveau, les répartitions selon l'âge et le sexe établies d'après les données des trois sources sont comparables. Cependant, l'enquête PERFORM produit un rapport de masculinité nettement plus faible pour le groupe des 30 à 49 ans (820) et légèrement plus élevé pour le groupe des 50 à 64 ans (993) que la NFHS (941 et 960, respectivement). Nous pensons que ces écarts sont dus, en partie, au fait que les travailleurs de terrain d'un des organismes chargés de l'enquête ont «poussé» les femmes à la fin de la période de procréation hors de cette tranche d'âge pour ne pas être obligés de remplir le calendrier des grossesses et les sections réservées aux antécédents du questionnaire. (Après avoir effectué une enquête supplémentaire, nous avons constaté que le rapport de masculinité pour la tranche des femmes de 50 à 64 ans était uniformément plus élevé dans les sept districts sous la responsabilité d'un organisme particulier que dans les autres.) Par conséquent, le nombre de femmes de 50 à 64 ans produit par l'enquête PERFORM est probablement un peu plus élevé qu'il ne l'est en réalité. Cela pourrait aussi signifier que les naissances attribuables à des femmes ayant effectivement moins de 50 ans ont été sous-dénombrées. Toutefois, comme il ne s'agit pas d'un groupe d'âge à haute fécondité, le biais n'est probablement pas très important.

3.2 Taille et caractéristiques des établissements

En se rendant dans les établissements sélectionnés par le biais des USÉ, ou grappes, et en y interviewant les membres du personnel, on peut produire un échantillon indépendant d'établissements de santé et de fournisseurs de services. (Sont inclus ceux qui fournissent des services de planification familiale à l'heure actuelle, ainsi que ceux susceptibles de le faire, c'est-à-dire les points de vente au détail (magasins de marchandises diverses, kirana et bureaux de prêteurs sur gage) inclus dans le nombre global estimé, mais qui ne distribuent pas de contraceptifs à l'heure actuelle.) Le dénombrement pondéré de ces points de fourniture de services figure au tableau 4. Le fait que nombre d'agents indépendants ne soient pas enregistrés, particulièrement les médecins «non qualifiés» (ou charlatans), rend plus difficile la validation des estimations de leur nombre. Selon Narayana, Cross et Brown (1994: tableau 8), en 1991, l'Uttar Pradesh comptait en tout

Tableau 3
Répartition en pourcentage de la population de Jure, selon l'âge et le sexe, d'après le SRS, la NFHS et l'Enquête PEFORM, pour la période de 1991 à 1995

Âge	SRS (1991)		NFHS (1992-93)			PERFORM (1995)		
	Hommes	Femmes	Hommes	Femmes	Rapport de masculinité	Hommes	Femmes	Rapport de masculinité
0-4	14,4	14,4	14,6	14,6	917	13,8	14,0	909
5-14	24,9	24,4	27,5	26,0	868	27,2	26,3	861
15-29	28,4	26,8	25,1	26,4	967	25,4	27,7	972
30-49	20,7	21,9	19,2	19,7	941	19,8	18,3	820
50-64	8,2	8,5	8,4	8,8	960	8,6	9,6	993
65+	3,6	4,0	5,2	4,4	718	5,2	4,1	702
Total	100,0	100,0	100,0	100,0		100,0	100,0	

Source des données du Sample Registration System (SRS): Bureau général de l'état civil de l'Inde (1993a)
Source des données de la NFHS: National Family Health Survey, Uttar Pradesh (1992-1993).

Tableau 1
Couverture des unités d'échantillonnage de l'Enquête PERFORM, Uttar Pradesh, 1995

Couverture de l'échantillon	Unités d'échantillonnage						
	Villages	îlots urbains	Ménages	Femmes admissibles	PFS fixes	Personnel des PFSF	Agents individuels
Nombre échantillonné	1 539	738	42 006	48 009	2 549	7 026	23 364
Nombre interviewé	1 539	738	40 633	45 277	2 428	6 320	22. 335
Taux de réponse	100,0	100,0	96,7	94,3	95,3	89,9	95,6

Nota: Les villages et les îlots urbains ont servi d'unités primaires d'échantillonnage; pour être admissible, les femmes devaient être couramment mariées et avoir entre 13 et 49 ans.
PFS = point de fourniture de services.

de réponse est très élevé pour les unités d'échantillonnage qui ont nécessité une interview sur place – variant de 94.3 % pour les femmes admissibles à 96.7 % pour les ménages. Pour les établissements de santé et les prestateurs individuels de services, le taux de réponse se chiffre à 95 %. Le taux n'est plus faible que pour les membres du personnel des établissements fixes. Toutefois, à 90 %, s'il n'est pas remarquable, il est quand même respectable. (Un type de membre du personnel, à savoir les infirmières auxiliaires – sages femmes, postées dans les sous-centres a été difficile à rejoindre, même après les trois essais habituels.)

3.1 Taille et caractéristiques de la population

Le tableau 2 permet de comparer, à l'échelle de la population, les valeurs de certains indicateurs démographiques obtenues d'autres sources à celles fournies par l'enquête PERFORM. Les chiffres indiquent que les résultats de l'enquête PERFORM concordent avec ceux du recensement, ainsi qu'avec ceux de la dernière National Family Health Survey (NFHS) effectuée dans l'État d'Uttar Pradesh à la fin de 1992 et au début de 1993 auprès d'un échantillon de 11 438 femmes de 13 à 49 ans ayant déjà été mariées. La population recensée a augmenté presque de 10,5 millions de personnes depuis le Recensement de 1991 et le pourcentage de ménages dans les régions urbaines est à peu près le même selon les trois sources. Le ratio du nombre de femmes au nombre d'hommes (rapport de masculinité) est légèrement plus faible dans le cas de l'enquête PERFORM (891) que dans celui de la NFHS (917). La comparaison des pourcentages de population

Tableau 2
Indicateurs démographiques de base pour l'Uttar Pradesh (Inde)

Indice	Uttar Pradesh				
	Recensement (1991)	NFHS (1992-93)	PERFORM (1995)	Erreur-type	Effet de plan
Population	139 112 287	*nd*	149 758 641	1 542 952	–
Pourcentage de population urbaine	19,8	22,6[a]	21,6[a]	0,6553	12,6095
Rapport de masculinité[b]	879	917	891	34,1010	0,9727
Pourcentage de 0 à 14 ans	39,1	41,8	40,2	0,1306	1,9049
Pourcentage de 65 ans et plus	3,8	4,8	4,7	0,0513	1,5789
Pourcentage appartenant à une caste désignée	21,0	18,0[a]	20,0[a]	0,3790	3,6536
Pourcentage appartenant à une tribu désignée	0,2	1,1[a]	3,1[a]	0,1818	4,4694
Pourcentage sachant lire et écrire[c]					
Hommes	55,7	65,3	67,6	0,3352	6,4634
Femmes	25,3	31,4	37,4	0,3824	8,6821
Total	41,6	49,9	53,3	0,3352	12,2385
Indice synthétique de fécondité	5,1	4,8	4,5	–	–
Prévalence des méthodes modernes de contraception	*nd*	18,5[d]	22,0[d]	0,3499	3,4111

[nd] Non disponible
[a] Calculé d'après le nombre de ménages
[b] Nombre de femmes pour mille hommes
[c] Calculé d'après la population de 7 ans et plus dans le cas du recensement et d'après la population de 6 ans et plus dans le cas de la NFHS et de l'Enquête PERFORM
[d] Pourcentage de femmes actuellement mariées de 15 à 49 ans utilisant une méthode moderne de contraception.

1. tous les établissements de santé privés et publics dans les USÉ rurales et urbaines sélectionnées;
2. tous les sous-centres, les centres primaires de santé, les centres communautaires de santé et les centres de soins post-partum qui fournissent des services à la population des USÉ rurales sélectionnées;
3. tous les hôpitaux privés comptant au moins 10 lits dans la ville la plus proche (dont la population est inférieure à 100 000 habitants) dans un rayon de 30 kilomètres des USÉ rurales sélectionnées;
4. tous les hôpitaux municipaux, les hôpitaux de district et les hôpitaux universitaires;
5. toutes les cliniques et tous les hôpitaux exploités par des organismes bénévoles, le secteur des soins organisés et les coopératives;
6. tous les PIS dans les villages et les îlots sélectionnés.

Il serait probablement utile de commencer par décrire la prestation organisée de soins de santé par le secteur public. Les résidents de tous les villages ont droit à obtenir des soins de santé auprès d'un sous-centre public (SC), d'un centre primaire de santé (CPS) ou d'un centre communautaire de santé (CCS). Les villages de 5 500 habitants et plus comptent souvent un sous-centre sur leur territoire. Environ six SC dépendent d'un CPS; à leur tour, les CPS sont rattachés à un CCS. Comme le CPS est parfois intégré au CCS, nous avons dû estimer le nombre combiné de CCS et de CPS, tout en estimant le nombre de SC séparément. (La croissance de la population a obligé à établir des «CPS supplémentaires» et à rerépartir en districts les zones desservies par les CPS originaux. Ces CPS supplémentaires sont inclus dans l'estimation du nombre de CPS.) On a effectué une visite sur place dans tous les SC attribués à un village échantillonné, ainsi qu'aux CPS et CCS affiliés.

Au moment de l'établissement de la liste et du relevé cartographique des ménages dans chaque îlot ou village, on a également dressé la liste et fait le relevé cartographique des PFSF et des PIS. De surcroît, dans chaque USÉ, on a interviewé des informateurs clés afin de prendre connaissance des points de fourniture de services de santé dont l'existence est moins manifeste. La sélection des points de fourniture de services – PFSF et PIS situés dans les limites des USÉ ou affiliés à un sous-centre de santé public – a été faite par recensement complet. Seuls les hôpitaux municipaux, les hôpitaux de district et les hôpitaux universitaires font exception et on leur a attribué un poids unitaire. Les probabilités de sélection des autres PFSF et PIS dépendent alors de la probabilité de sélection de l'USÉ et l'inverse de cette dernière représente le poids du PFSF ou du PIS. On a calculé les poids appliqués aux CCS, aux CPS et aux SC selon la méthode décrite plus bas, après avoir décelé certaines «défaillances» sur le terrain lors de la sélection de ce type d'établissements. (On discutera de ces défaillances plus tard.)

Comme les CCS et les CPS sont associés à plus d'une USÉ, nous avons supposé qu'il existe un CPS pour 30 000 habitants (chiffre qui représente à peu près la moyenne réelle pour l'État d'Uttar Pradesh) et qu'un SC dessert environ 5 500 personnes (les chiffres moyens réels pour les districts varient de 4 000 à 6 500). Dans ces conditions, le poids appliqué aux CCS/CPS pour chaque USÉ sélectionnée est

$$W_{\mathrm{CCS/CPS}} = \frac{\text{Population totale de l'USÉ sélectionnée}}{30\ 000} * VW_{lijk} \,(\text{ou } UW_{1ijk})$$

et le poids appliqué aux SC pour chaque USÉ sélectionnée est

$$W_{SC} = \frac{\text{Population totale de l'USÉ sélectionnée}}{5\ 500} * VW_{1ijk} \,(\text{ou } UW_{1ijk}).$$

Il a fallu corriger les poids calculés pour les PFSF non autosélectionnés afin de tenir compte de la multiplicité, c.-à-d. les situations où un PFSF est sélectionné dans l'échantillon en rapport avec plus d'une USÉ. Par exemple, il arrive qu'un CCS/CPS soit sélectionné à cause de deux USÉ. Le cas échéant, on a appliqué au CCS/CPS, un poids égal à la somme des poids des deux USÉ choisies, c.-à-d. $W_{\mathrm{CCS/CPS}}$.

2.3 Mise en oeuvre de l'enquête

Le travail sur le terrain de l'enquête PERFORM a été effectué de juin à septembre 1995 dans l'État d'Uttar Pradesh. L'enquête a été exécutée sous contrat par quatre organismes choisis selon une méthode d'approvisionnement concurrentiel. Un organisme qui avait testé le plan de l'enquête PERFORM dans un district l'année auparavant a joué le rôle d'organisme nodal ou coordonnateur. Les coordonnateurs et le superviseur du projet ont reçu une formation d'instructeur principal, y compris la participation à un essai préliminaire sur le terrain. L'enquête PERFORM proprement dite a été effectuée par des équipes de six personnes comprenant un superviseur, une vérificatrice, un intervieweur et quatre intervieweuses. Chaque organisme chargé du travail sur le terrain a engagé, en moyenne, trois équipes pour couvrir un district, soit 18 employés régionaux en tout pour la collecte des données par district (ou 21 équipes comptant en tout 126 employés régionaux pour couvrir 7 districts). La supervision globale sur le terrain a été confiée à une équipe de quatre personnes désignées spécialement, assignées chacune à un des organismes chargés de l'exécution de l'enquête. Après vérification sur le terrain, les questionnaires ont été acheminés au bureau central des organismes chargés de l'enquête aux fins de la saisie et de l'épuration des données.

3. RÉSULTATS

Le tableau 1 donne la couverture de l'échantillon de l'enquête PERFORM en ce qui concerne le nombre d'unités de chaque type sélectionnées, le nombre d'unités effectivement interviewées et le taux de réponse. Le taux

$$r_k = 2 * \frac{m_k}{M}$$

où M représente la population totale de la division ($M = \sum_{k=1}^{t} m_k$) et où t représente le nombre total de districts dans la division.

2.2.2 Probabilité de sélection des villages et des ménages

Représentons par n_{ijk} le nombre de ménages dans le i-ième village, la j-ième strate et le k-ième district. Alors, p_{ijk}, c'est-à-dire la probabilité de sélectionner le village i dans la j-ième strate et le k-ième district est donnée par

$$p_{ijk} = a_{jk} * \frac{n_{ijk}}{N_{jk}} * r_k$$

où a_{jk} et N_{jk} représentent, respectivement, le nombre de villages sélectionnés et le nombre total de ménages dans la j-ième strate et le k-ième district.

Représentons par q_{ijk} la probabilité de sélectionner un ménage dans les régions rurales d'un district sélectionné. Alors, on peut calculer q_{ijk} selon l'équation

$$q_{ijk} = p_{ijk} * \frac{20}{n_{ijk}}$$

où 20 est le nombre de ménages tirés du village sélectionné.

Les poids appliqués aux villages et aux ménages sont alors égaux à l'inverse de la probabilité de sélection de ces derniers, c.-à-d. $1/p_{ijk}$ et $1/q_{ijk}$, et sont représentés par VW_{1ijk} et HW_{1ijk}, respectivement.

2.2.3 Probabilité de sélection des villes, des îlots urbains et des ménages

La probabilité de sélectionner de la j-ième ville dans le k-ième district, t_{jk}, est égale à

$$t_{jk} = 1 \text{ si la population de la ville est} > 100\,000$$

$$t_{jk} = c_k \frac{s_{jk}}{S_k} \text{ si la population de la ville est} < 100\,000$$

où s_{jk} représente le nombre total de ménages dans la j-ième ville (ayant une population <100 000) du k-ième district, c_k représente le nombre de villes sélectionnées dans le district k et S_k représente le nombre total de ménages dans les villes dont la population est inférieure à 100 000 dans le district k.

Représentons par u_{ijk} la probabilité de sélectionner le i-ième îlot dans la j-ième ville et le k-ième district. Alors, u_{ijk} est donnée par

$$u_{ijk} = b_{jk} * \frac{x_{ijk}}{Y_{jk}} * t_{jk} * r_k$$

où b_{jk} représente le nombre d'îlots urbains sélectionnés et où Y_{jk} représente le nombre total de ménages dans la j-ième ville du k-ième district, et x_{ijk} représente le nombre de ménages dans le i-ième îlot de la j-ième ville du k-ième district.

La probabilité de sélectionner un ménage de l'i-ième îlot et du k-ième district, représentée par v_{ijk}, est donné par

$$v_{ijk} = u_{ijk} * \frac{15}{x_{ijk}}$$

où 15 est le nombre de ménages tirés de l'îlot urbain sélectionné.

Les poids appliqués aux îlots urbains et aux ménages sont alors égaux à l'inverse de la probabilité de sélection de ces îlots ou ménages, c.-à-d. $1/u_{ijk}$ et $1/v_{ijk}$, et sont représentés par UW_{1ijk} et HW_{1ijk}, respectivement. Puisqu'au niveau de la population, les estimations sont fondées sur des personnes, on a appliqué à tous les membres d'un même ménage sélectionné le poids attribué à ce ménage. Aucune méthode de sélection n'a été appliquée aux membres d'un ménage admissibles comme répondants.

2.2.4 Correction pour la non-réponse au questionnaire sur le ménage et pour le suréchantillonnage des îlots urbains

Pour tenir compte de la non-réponse dans le calcul des poids appliqués aux ménages, on suppose que la non-réponse est aléatoire dans le village (ou dans l'îlot) et on procède comme suit:

Posons que n_1 est le nombre de ménages sélectionnés et que n_2 est le nombre de ménages où sont effectuées des interviews. Alors, le poids corrigé en fonction de la non-réponse qu'on attribue aux ménages est défini comme

$$HW_{2ijk} = HW_{1ijk} * \frac{n_1}{n_2}.$$

Le poids final appliqué aux ménages comprend aussi une correction de la proportion de population urbaine dans le district, dans les cas où il y a eu suréchantillonnage des îlots urbains (districts dont la population urbaine est inférieure à 20 %).

Posons que n_3 est la proportion réelle de population urbaine dans un district et que n_4 est la proportion de population urbaine dans l'échantillon. Alors, le poids corrigé pour tenir compte de la non-réponse et du suréchantillonnage des îlots appliqué aux ménages est défini par

$$HW_{3ijk} = HW_{2ijk} * \frac{n_3}{n_4}.$$

2.2.5 Sélection des points de fourniture de services dans les échantillons de district

Pour obtenir un échantillon probabiliste des points de fourniture de services, on a sélectionné les PFSF et les PIS en rapport avec les USÉ, c.-à-d. les villages ou les îlots, de la façon suivante:

d'estimations pour les principaux indicateurs de niveau de population. Une taille globale cible d'échantillon de 1 627 femmes de 13 à 49 ans ayant déjà été mariées a été nécessaire pour déceler une variation de cinq points de la prévalence de la contraception (avec $\alpha = 0.05$ et $1 - \beta = 0.90$) au niveau du district. Comme on s'attend à ce que le nombre par ménage de femmes de 13 à 49 ans ayant déjà été mariées soit de 1,15, on obtiendrait le nombre requis de femmes déjà mariées en rendant visite à un échantillon de 1 415 ménages. En se donnant une marge de sécurité supplémentaire de 5 % pour tenir compte de la non-réponse et de la non-disponibilité, on a estimé qu'un échantillon cible de 1 725 femmes de 13 à 49 ans ayant déjà été mariées tiré de 1 500 ménages serait suffisant. Le diagramme schématique du plan d'échantillonnage est présenté à la figure 1.

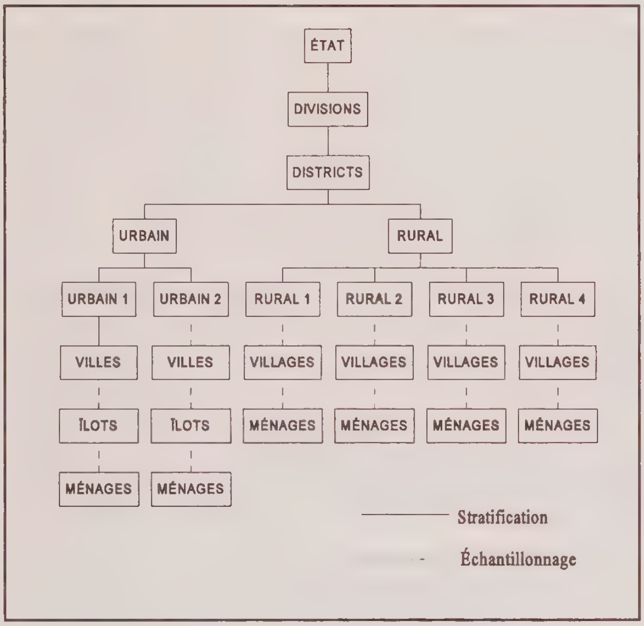


Figure 1. Diagramme schématique du plan d'échantillonnage PERFORM

De surcroît, on a stratifié les districts en régions rurales et urbaines. Selon les définitions du Recensement de l'Inde, tous les lieux comptant une municipalité, une «corporation» municipale, un conseil de canton ou un comité régional notifié, ainsi que tous les autres lieux comptant au moins 5 000 habitants dont au moins 75 % de la population active masculine effectue des travaux non agricoles et dont la densité de population est au moins égale à 400 personnes par kilomètre carré sont classés dans la catégorie des régions urbaines. Les îlots urbains et les villages ruraux servent d'unités secondaires d'échantillonnage (USÉ). Les 1 500 ménages à échantillonner dans chaque district ont été répartis entre les régions rurales et urbaines proportionnellement à la taille de la population du district. Cependant, dans les cas où la proportion allouée de population urbaine était inférieure à 20 %, on a fixé l'allocation de ménages dans la région urbaine à 20 %, afin d'être certain de couvrir un nombre suffisant de points de fourniture de services de santé.

Dans les régions rurales, on a sélectionné les ménages selon un plan d'échantillonnage stratifié à deux degrés. On a d'abord réparti les villages des régions rurales en quatre strates, selon la taille de la population, de la façon suivante:

Strate	Taille de la population du village
I	100 - 499
II	500 - 1 999
III	2 000 - 4 999
IV	5 000 et plus.

On a exclu de la liste les villages comptant moins de 100 habitants ou moins de 20 ménages (pareils villages étaient rares dans le cas de l'étude décrite ici). Le nombre de villages à sélectionner dans chaque district a été réparti proportionnellement entre les quatre strates. Pour sélectionner les villages, on a commencé par les ordonner dans la strate selon le taux de d'alphabétisation des femmes, puis on a sélectionné le nombre requis de village par une méthode d'échantillonnage avec probabilité proportionnelle à la taille. Après avoir dressé la liste et fait le relevé cartographique de tous les ménages dans les villages sélectionnés, on a tiré un nombre cible de 20 ménages dans chaque village selon une méthode d'échantillonnage systématique. On a réparti les villages comptant plus de 500 ménages ou 2 500 habitants et plus (certains villages de la strate III et tous ceux de la strate IV) en quatre groupes et sélectionné de ces deux groupes pour l'établissement de la liste et la sélection des ménages. On a sélectionné les 20 ménages requis en tirant dix ménages de chaque groupe par échantillonnage aléatoire systématique.

Dans les régions urbaines, on a également sélectionné les ménages selon un plan d'échantillonnage stratifié à deux degrés. On a stratifié les villes des régions urbaines de chaque district d'après la taille de la population, de la façon suivante:

Strate	Taille de la population de la ville
I	100 000 et plus
II	Moins de 100 000.

On a sélectionné toutes les villes de la strate I avec certitude. Dans le cas de la strate II, on a ordonné les villes selon la taille de la population, puis on a sélectionné le nombre requis par échantillonnage avec probabilité proportionnelle à la taille. Ensuite, de chaque ville échantillonnée, on a échantillonné au moins deux îlots avec probabilité proportionnelle à la taille. Enfin, on a dressé la liste et fait le relevé cartographique de tous les ménages dans les îlots sélectionnés et on a tiré 15 ménages de chaque îlot par échantillonnage aléatoire systématique.

2.2.1 Probabilité de sélection des districts

Représentons par m_k la population du k-ième district dans une division. Comme on doit sélectionner deux districts dans chaque division, la probabilité de sélectionner le k-ième district d'une division r_k est donnée par

des méthodes d'enquête innovatrices permettant de fournir aux planificateurs et aux gestionnaires des services de santé le plus de renseignements possible en perdant le moins de précision possible.

Nous présentons ici les résultats d'une enquête par échantillonnage en grappes à plusieurs degrés conçue pour estimer la population et les caractéristiques des établissements de santé et des populations de clients visées. L'échantillon en grappes de l'enquête, qui a été effectuée dans le grand État d'Uttar Pradesh, en Inde du Nord, a servi de base pour la sélection des établissements de santé et des ménages. Puis, on a sélectionné les prestateurs de soins dans les établissements et les femmes mariées en âge de procréation dans les ménages. L'enquête a été conçue pour produire des échantillons indépendants d'établissements de santé, de membres du personnel, de ménage et de population de clients des services de santé.

Dans la section qui suit, nous décrivons le plan de sondage, son contenu et les méthodes de travail sur le terrain appliquées en Uttar Pradesh. Puis, à la section suivante, nous comparons les résultats obtenus pour les établissements de santé et pour la population de clients et, à la dernière section, nous dégageons de l'application de la méthode en Uttar Pradesh certaines leçons au chapitre de la conception d'enquêtes. Ces enseignements seront particulièrement importants au moment de la répétition de l'enquête prévue dans deux ans, mais ils sont aussi susceptibles d'intéresser d'autres pays qui voudraient adopter le plan d'échantillonnage en grappes enchaînées.

2. L'ENQUÊTE PERFORM EN UTTAR PRADESH

L'enquête PERFORM ou Project Evaluation Review For Organizational Resource Management (examen évaluatif des projets pour la gestion des ressources organisationnelles) a pour objectif d'évaluer des indicateurs de référence pour un grand projet de planification familiale, baptisé Innovations in Family Planning Services (IFPS) project exécuté au Uttar Pradesh et financé conjointement par le gouvernement de l'Inde et par la U.S. Agency for International Development. L'État d'Uttar Pradesh compte plus de 140 millions d'habitants et, pris individuellement, représenterait le cinquième plus grand pays en voie de développement.

2.1 Contenu

L'estimation d'indicateurs pour l'IFPS doit être effectuée à trois niveaux, à savoir 1) les points de fourniture de services (PFS) publics et privés, 2) les prestateurs de services faisant partie du personnel des PFS ou des établissements de santé et 3) la population de clientes, c'est-à-dire les femmes en âge de procréation. Comme l'IFPS a pour objectif d'améliorer l'environnement dans lequel a lieu la prestation de services de planification familiale, il est impératif de mesurer les indicateurs à ce niveau, mais de façon à ce que la mesure puisse être reliée aux femmes qui vivent dans cet environnement.

Par conséquent, l'équipe de l'enquête PERFORM a conçu sept questionnaires:

1-2) questionnaire visant un îlot urbain ou un village pour dresser la liste de tous les fournisseurs éventuels et réels de services de santé dans le village ou l'îlot échantillonné;

3) questionnaire visant les points de fourniture de services fixes (PFSF) pour recueillir des renseignements sur les membres du personnel, les services, l'équipement, les fournitures et les activités de formation et de motivation auprès des établissements publics et privés échantillonnés;

4) questionnaire s'adressant aux membres du personnel, à faire remplir par tous les membres du personnel des PFSF qui offrent des services de planification familiale (recensés d'après les réponses au questionnaire visant les PFSF) pour évaluer leurs compétences et leur expérience;

5) questionnaire s'adressant aux prestateurs individuels de services (PIS), à faire remplir par toutes les personnes travaillant en-dehors des établissements autonomes (PFSF) qui prodiguent actuellement ou qui pourraient prodiguer des services de planification familiale, dont les services de médecins particuliers, de pharmaciens, de sages-femmes, de travailleurs de la santé non spécialisés et de détaillants;

6) questionnaire visant les ménages, à faire remplir par les chefs des ménages échantillonnés pour recenser les membres du ménage et recueillir des données sur les caractéristiques démographiques et sociales;

7) questionnaire personnel s'adressant aux femmes mariées à l'heure actuelle, âgées de 13 à 49 ans (repérées grâce au questionnaire sur le ménage) pour collecter des renseignements sur ce qu'elles savent de l'existence de services de santé et sur l'utilisation passée, courante et prévue de ces services, sur les grossesses récentes et les comportements à l'égard de la contraception et sur d'autres caractéristiques générales.

2.2 Plan d'échantillonnage

L'enquête PERFORM a été conçue pour estimer les caractéristiques des établissements de santé et de leur population de clients au niveau de l'État, de la région, de la division et du district. Ce dernier est important, car il s'agit du niveau où est concentré le lancement de méthodes innovatrices et d'efforts supplémentaires dans le cadre de l'ISPS. Au moment de la conception de l'enquête, l'État d'Uttar Pradesh comptait 14 divisions administratives. Dans chacune de ces divisions, on a sélectionné deux districts par échantillonnage avec probabilité proportionnelle à la taille (PPT). Ces unités géographiques possèdent des limites politico-administratives, donc des services d'administration publique. En outre, on a agrégé les districts en cinq groupes régionaux.

On a fixé à 1 500 le nombre total de ménages à sélectionner dans chaque district. On a en effet déterminé qu'un échantillon de 1 500 ménages suffirait pour la production

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 147-157
Statistique Canada

Estimation de la population et des caractéristiques des établissements de santé et des populations de clients au moyen d'un plan d'échantillonnage à plusieurs degrés avec enchaînement

K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN et G. NARAYANA[1]

RÉSUMÉ

Le présent article montre l'utilité d'un plan de sondage à plusieurs degrés pour obtenir le dénombrement total des établissements de santé et de la population de clients éventuels dans une région. Le plan décrit a été utilisé pour effectuer une enquête à l'échelle de l'État d'Uttar Pradesh, en Inde, au milieu de 1995. Il comprend la sélection d'un échantillon aréolaire en grappes à plusieurs degrés où l'unité primaire d'échantillonnage est soit un îlot urbain, soit un village rural. On a fait le relevé cartographique, dressé la liste et sélectionné tous les points de fourniture de services de santé, qu'il s'agisse d'établissements autonomes ou d'agents de distribution, situés dans les unités primaires d'échantillonnage ou assignés officiellement à ces dernières. On a tiré un échantillon systématique de ménages et interviewé toutes les femmes faisant partie de ces ménages qui satisfaisaient les critères prédéterminés d'admissibilité. On a appliqué des poids d'échantillonnage aux établissements ainsi qu'aux personnes. Pour les établissements, les poids sont corrigés pour tenir compte du fait que certains établissements desservent plusieurs unités secondaires d'échantillonnage. Pour les personnes, on a corrigé les poids pour tenir compte des taux de réponse à l'enquête. L'estimation par sondage du nombre total d'établissements publics concorde bien avec les totaux publiés. Pareillement, l'estimation de la population de clientes calculée d'après l'enquête concorde avec le chiffre total du Recensement de 1991.

MOTS CLÉS: Enquête par sondage; évaluation des programmes; services de santé; pays en voie de développement.

1. INTRODUCTION

Pour évaluer l'incidence des programmes de services de santé sur la santé de la population, il est souvent nécessaire de connaître le nombre et les caractéristiques des établissements de santé et des clients éventuels. Or, pareils renseignements font souvent défaut dans les pays en voie de développement où les dossiers sur les programmes et les systèmes d'enregistrement des données de l'état civil sont en général incomplets et mal tenus à jour.

Pour obtenir des renseignements courants sur l'état de santé, l'utilisation des services de santé, le rendement des services et les besoins des clients, les responsables des programmes s'appuient sur des enquêtes par sondage occasionnelles, souvent conçues et effectuées indépendamment les unes des autres, à un niveau infrarégional (Aday 1991; Ross et McNamara 1983). Néanmoins, certaines enquêtes sur la démographie et sur la santé (Macro International 1996) fournissent un profil national de divers aspect de la santé de la population, comme la fécondité, la mortalité infantile et le bien-être nutritionnel. L'avantage distinct que présente un échantillon national de population pour la planification des programmes de santé tient au fait qu'il permet d'évaluer les attitudes et les comportements des clients ainsi que des non-clients. Les statistiques sur les services offerts par les programmes se limitent aux clients réels et ne permettent pas toujours de brosser le tableau le plus à jour qui soit de l'utilisation des services.

Outre le comportement des clients, il est utile de surveiller l'offre de services ainsi que la qualité de ceux-ci, mais cet exercice nécessite un examen distinct de la fourniture de services par les établissements de santé ou par les établissements connexes. Les efforts déployés à cet égard dans les pays en voie de développement, comme les études d'analyse de la situation (Miller, Ndhiovu, Gachara et Fisher 1991), incluent l'exécution, auprès des établissements de santé, d'enquêtes probabilistes qui donnent un aperçu national du rendement des programmes. Cependant, ces enquêtes probabilistes sont souvent limitées à l'examen des programmes de santé publique. En effet, l'enregistrement incomplet ou inexact des fournisseurs de services de santé du secteur privé, comme les cliniques privées ou les pharmacies, empêche de recourir à cette méthode d'enquête pour suivre les tendances de la prestation des soins de santé par ce secteur.

Les ressources dont on dispose pour étendre et améliorer la fourniture de services de santé sont de plus en plus limitées tant dans les pays en voie de développement que dans les pays développés. Par conséquent, toutes les parties concernées cherchent à mieux utiliser les ressources existantes pour effectuer le suivi et l'évaluation, particulièrement au moyen d'enquêtes. On devrait donc élaborer

[1] Kaushalendra K. Singh, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Statistics, Faculty of Science, Banaras Hindu University, Varanasi 221005 India; Amy O. Tsui, Director, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Maternal and Child Health, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Chirayath M. Suchindran, Carolina Population Center, University of North Carolina at Chapel Hill, CB #8120 University Square, Chapel Hill, NC 27516-3997 and Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, CB #7400 Rosenau Hall, Chapel Hill, NC 27599-7400; Gaade Narayana, The Futures Group International, 1050 17th Street, N.W., Suite 1000, Washington, DC 20036.

LITTLE, R.J.A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

LITTLE, T.C. (1996). Models for nonresponse adjustment in sample surveys. Thèse de doctorat, Department of Statistics, University of California, Berkeley.

LITTLE, T.C., et GELMAN, A. (1996). A model for differential nonresponse in sample surveys. Rapport technique.

LONGFORD, N.T. (1996). Small-area estimation using adjustment by covariates. *Qüestió* 20, à paraître.

NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

VOSS, D.S., GELMAN, A., et KING, G. (1995). Pre-election survey methodology: details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59, 98-132.

$$E(t(\gamma_l) \mid y, \tau^{\text{vieux}}) =$$

$$\| E(\gamma_l \mid y, \tau^{\text{vieux}}) \|^2 + \text{trace}(\text{var}(\gamma_l \mid y, \tau^{\text{vieux}})).$$

Puisqu'on ne peut traiter analytiquement ces deux termes dans le cas de notre modèle, nous utilisons les approximations suivantes que l'on peut obtenir facilement: (1) on s'approche de $E(\gamma_l \mid y, \tau^{\text{vieux}})$ avec une estimation $\hat{\gamma}_l$, fondée sur y et l'estimation τ^{vieux}, et (2) on s'approche de $\text{var}(\gamma_l \mid y, \tau^{\text{vieux}})$ de la courbure de la log-vraisemblance avec l'estimation $\hat{V}_{\gamma_l} = (-L''(\hat{\gamma}_l))^{-1}$. Nous mettons ces approximations à jour itérativement pour tous les $l = 1, ..., L$ simultanément, pour converger vers une estimation du maximum de vraisemblance approximative $(\hat{\tau}_1, ..., \hat{\tau}_L)$. Étant donné une valeur provisoire initiale de τ^{vieux}, l'algorithme procède vers la convergence par itération des deux étapes suivantes.

Étape *E* approximative. Résoudre les équations de vraisemblance itérativement, comme décrit ci-après. Se servir des estimations $\hat{\beta}$ pour obtenir une approximation de $E(t(\gamma_l) \mid y, \tau^{\text{vieux}})$ pour chaque $l = 1, ..., L$.

Nous résolvons les équations de vraisemblance $d/d\beta\, L(\beta \mid y, \tau) = 0$ au moyen de moindres carrés pondérés itérativement, en incluant une approximation normale de la vraisemblance $p(y \mid \beta) = \Pi_i\, p(y_i \mid \beta)$, fondée sur l'approximation locale du modèle de régression logistique par un modèle de régression linéaire (voir Gelman et coll. 1995, p. 391). Posons que $\eta_i = (Z\beta)_i$ est le prédicteur linéaire de la i-ième observation. En commençant par la valeur provisoire courante de $\hat{\beta}$, posons que $\hat{\eta} = Z\hat{\beta}$. Alors, une extension de la série de Taylor à $L(y_i \mid \eta_i)$ donne $z_i \approx N(\eta_i, \sigma_i^2)$, où

$$z_i = \hat{\eta}_i + \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \left(y_i - \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} \right)$$

$$\sigma_i^2 = \frac{(1 + \exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)}.$$

Représentons par $\hat{\Sigma}_\beta$ la valeur de Σ_β fondée sur l'introduction de l'estimation courante $\hat{\tau}$ et posons que $\hat{\Sigma}_z = \text{diag}(\sigma_i^2)$. Alors, nous obtenons une estimation et une matrice de variance à jour en nous servant des moindres carrés pondérés fondés sur la distribution normale a priori et sur l'application de l'approximation normale à la vraisemblance de régression logistique:

$$\hat{\beta} = (Z' \hat{\Sigma}_z^{-1} Z + \hat{\Sigma}_\beta^{-1})^{-1} Z' \hat{\Sigma}_z^{-1} z \qquad (2)$$

$$\hat{V}_\beta = (Z' \hat{\Sigma}_z^{-1} Z + \hat{\Sigma}_\beta^{-1})^{-1}. \qquad (3)$$

Nous poursuivons l'itération jusqu'à la convergence, puis nous utilisons $\hat{\beta}$ et les éléments appropriés de $\hat{V}_\beta$ pour estimer $\text{var}(\gamma_l \mid y, \tau^{\text{vieux}})$.

Étape *M*. Maximiser sur les paramètres τ_1 pour obtenir $\tau_l^{\text{nouveau}} = (\hat{E}(t(\gamma_l) \mid y, \tau^{\text{vieux}})/K_l)^{1/2}$ pour chaque $l = 1, ..., L$. Remplacer la valeur de τ^{vieux} par celle de τ^{nouveau} et retourner à l'étape *E* approximative.

Une fois que l'algorithme EM a convergé vers une estimation $\hat{\tau}$, nous tirons β d'une approximation normale de la distribution conditionnelle a posteriori $p(\beta \mid y, \hat{\tau})$, en nous servant des valeurs produites par les équations (2) et (3) à la dernière étape EM comme matrice de la moyenne et de la variance dans l'approximation normale. Pour chaque tirage du paramètre vectoriel β, nous calculons les moyennes des catégories, $\pi = \text{logit}^{-1}(X\beta)$, et tous les totaux de population que l'on veut étudier, en comptant N_j unités de population dans chaque catégorie j.

BIBLIOGRAPHIE

BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., et ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (avec discussion). *Journal of the American Statistical Association* 88, 1149-1166.

CLAYTON, D.G. (1996). Generalized linear mixed models. In *Practical Markov Chain Monte Carlo*, Éds. W. Gilks, S. Richardson et D. Spiegelhalter, 275-301. New York: Chapman & Hall.

DEMING, W., et STEPHAN, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals of Mathematical Statistics* 11, 427-444.

DEMPSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (avec discussion). *Journal of the Royal Statistical Society*, 39, 1-38.

GELMAN, A., CARLIN, J.B., STERN, H.S., et RUBIN, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

GELMAN, A., et KING, G. (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23, 409-451.

HOLT, D., et SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society*, 142, 33-46.

KRIEGER, A.M., et PFEFFERMANN, D. (1992). Estimation par la méthode du maximum de vraisemblance dans des enquêtes par sondage complexes. *Techniques d'enquête*, 18, 241-256.

LAZZERONI, L.C., et LITTLE, R.J.A. (1997). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, à paraître.

LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

La stratification a posteriori bayésienne est surtout utile pour calculer des estimations sur des sous-ensembles de la population (p. ex., États distincts dans le cas des sondages d'opinion aux États-Unis) pour lesquelles l'effectif de l'échantillon est faible. Un domaine connexe dans lequel la modélisation devrait donner de bons résultats est celui du regroupement d'enquêtes effectuées par des organismes distincts, avec modélisation subordonnée à toutes les variables susceptibles d'avoir une influence sur la non-réponse dans l'une ou l'autre enquête. De surcroît, les méthodes décrites dans le présent article peuvent manifestement être appliquées à des réponses continues en remplaçant les modèles de régression logistique par d'autres modèles linéaires généralisés.

Dans le cas de la modélisation bayésienne, notre objectif ne consiste pas à ajuster un modèle subjectivement «vrai» aux données ni aux réponses sous-jacentes, mais plutôt à estimer avec une précision raisonnable la réponse moyenne, en imposant comme contrainte un grand ensemble de covariables observées complètement. Des modèles plus précis des réponses devraient permettre de faire des inférences plus exactes – néanmoins, même le simple modèle à effets mixtes échangeables que nous avons ajusté, avec hyperparamètres estimés d'après les données, devrait donner de meilleurs résultats que les valeurs extrêmes produites par les modèles à effets constants ou par l'adoption d'une valeur nulle pour les coefficients. En dernière analyse, l'objectif de la modélisation probabiliste et de l'inférence bayésienne dans le contexte d'une enquête par sondage est de pouvoir se servir de la profusion de renseignements au niveau des strates a posteriori (p. ex., données de recensement classées selon le sexe, le groupe ethnique, l'âge, le niveau de scolarité et l'État) pour rajuster les données d'une enquête effectuée sur un échantillon relativement petit.

Les méthodes de modélisation que nous proposons pourraient poser diverses difficultés. Si on adapte à un grand nombre de catégories un modèle trop faible (comme le modèle avec effets d'État non lissés), la variabilité des estimations résultantes pourrait être trop forte. Si on ne connaît pas la répartition de la population pour les variables utilisées pour effectuer la stratification a posteriori (par exemple, rajustement pour une variable qui n'est pas mesurée ou qui est mesurée de façon imprécise au moment du recensement), alors il faut modéliser les N_j également, ce qui donne un surcroît de travail. Naturellement, l'application de la méthode itérative du quotient à ces variables nécessiterait aussi du travail supplémentaire. Puisque toutes les méthodes, y compris la méthode itérative du quotient et les méthodes de régression, se fondent sur la supposition qu'on peut ne pas tenir compte de la non-réponse, les inférences produites seront incorrectes si les variables non mesurées ont une incidence sur la non-réponse et sont corrélées aux résultats que l'on veut étudier.

Les méthodes décrites ici visent à améliorer les corrections par stratification a posteriori du genre de la méthode itérative du quotient et ne sont pas destinées, en soi, à apporter une correction pour la non-réponse dont on doit tenir compte. Cependant, en permettant de faire le rajustement pour un plus grand nombre de variables, la stratification a posteriori bayésienne devrait rendre possible l'utilisation de modèles pour lesquels l'hypothèse selon laquelle on ne doit pas tenir compte de la non-réponse est plus raisonnable. Considérer un grand nombre de catégories pour la stratification a posteriori (p. ex., dans 48 États) crée des problèmes quand on applique les méthodes classiques de pondération, car nombre de catégories ne comptent que quelques répondants, voire aucun. Cependant, il est intéressant de noter que le fait de travailler avec un grand nombre de catégories rend parfois la modélisation bayésienne plus fiable: un plus grand nombre de catégories signifie un plus grand nombre d'effets aléatoires dans la régression, situation susceptible de faciliter l'estimation des composantes de la variance.

REMERCIEMENTS

Nous remercions Xiao-Il Meng et plusieurs évaluateurs de leurs commentaires précieux, ainsi que la *National Science Foundation* pour la bourse DMS-9404305 et le *Young Investigator Award* DMS- 9457824.

ANNEXE: CALCUL

Nous utilisons un algorithme de type EM pour estimer les hyperparamètres τ_l. Étant donné ces paramètres, nous tirons l'échantillon à partir de la distribution a posteriori des coefficients β au moyen d'une approximation normale de la vraisemblance de régression logistique. Nous utilisons cette approximation en raison de sa simplicité et parce qu'elle est réaliste pour des enquêtes relativement importantes, comme celles de l'application que nous décrivons à la section 3. Au besoin, des calculs plus précis peuvent être effectués au moyen de l'échantillonneur de Gibbs et de l'algorithme de Metropolis (consulter Clayton 1996), peut-être en utilisant l'algorithme décrit ici comme point de départ.

Si la distribution des données est normale et que les moyennes sont linéaires dans les coefficients de régression, on peut utiliser l'algorithme EM pour estimer les composantes de la variance (Dempster, Laird et Rubin 1977) en traitant le vecteur des coefficients β comme des «données manquantes». Dans ce contexte, la log-vraisemblance d'avoir des «données complètes» pour τ_l est

$$L(\tau_l \mid \gamma_l) = \text{const} - K_l \log \tau_l - \frac{1}{2\tau_l^2} \sum_{k=1}^{K_l} \gamma_{kl}^2,$$

de sorte que la statistique exhaustive pour τ_l est $t(\gamma_l) = \sum_{k=1}^{K_l} \gamma_{kl}^2$. Étant donné l'estimation courante τ^{vieux}, l'espérance de la statistique exhaustive est

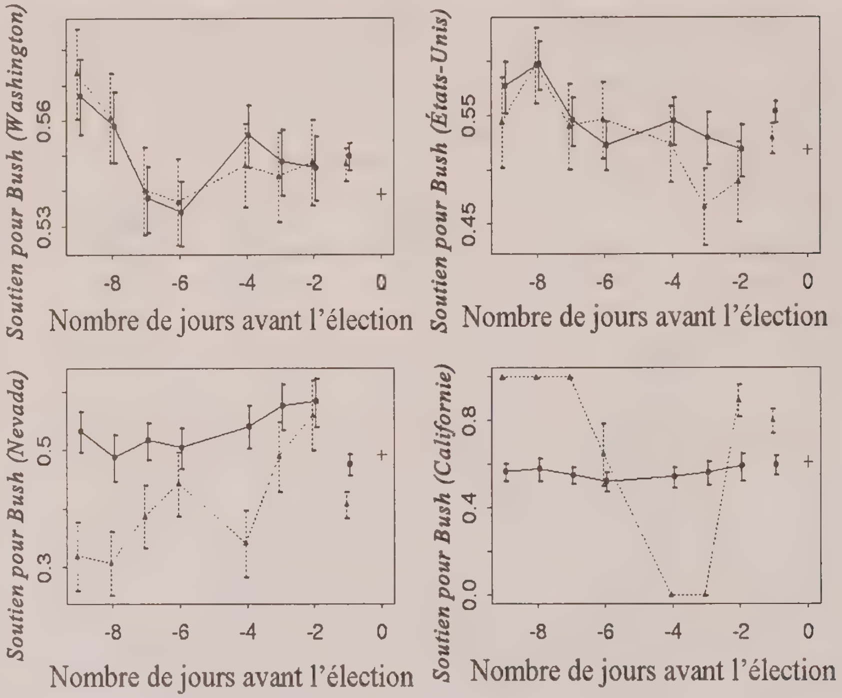


Figure 3. Soutien pour Bush estimé séparément d'après sept sondages d'opinion distincts exécutés peu de temps avant l'élection pour a) l'ensemble des États-Unis (sauf l'Alaska, Hawaii et le district de Columbia), b) un grand État (Californie), c) un État de taille moyenne (Washington) et d' un petit État (Nevada). Dans chaque graphique, la ligne en pointillés représente les estimations par la méthode itérative du quotient et la courbe en trait plein, celle produite par le modèle hiérarchique, et les barres d'erreur indiquent les limites de confiance de 50 % pour la méthode du quotient et les intervalles a posteriori de 50 % pour les estimations fondées sur le modèle. Les sondages d'opinion ont eu lieu entre le neuvième et le deuxième jours précédant l'élection. Les estimations fondées sur les données agrégées des sondages sont indiquées au temps «– 1», et les résultats réels de l'élection sont indiqués au temps «0» dans chaque graphique.

le prévoirait en neutralisant simplement les covariables covariables démographiques (cette prévision serait l'estimation pour l'État de Washington calculée d'après le modèle où la valeur des effets d'État est maintenue nulle, prévision qui, d'après le tableau 1, est égale à 0,58). Néanmoins, aucun sondage, pris isolément, ne fournissait suffisamment de données pour soutenir, de façon convaincante, que l'opinion dans l'État de Washington s'écartait à ce point de la moyenne nationale. Par conséquent, l'estimation bayésienne a rétréci plus fortement les estimations tirées de ces sondages. S'il paraît étrange a priori, ce comportement n'en est pas moins approprié: dans le cas d'une enquête de petite envergure, on dispose de moins de renseignements sur les diverses catégories résultant de la stratification a posteriori et les estimations axées sur le modèle produisent, pour chacune de ces catégories, une estimation plus proche de la moyenne d'échantillon. Quand on agrège les résultats des sept sondages, on dispose de plus de renseignements et le modèle s'appuie davantage sur les données dans chaque catégorie. C'est essentiellement par ce procédé que la méthode de Bayes contrebalance les difficultés que pose la stratification a posteriori comptant un nombre trop petit ou trop grand de catégories.

4. DISCUSSION

La stratification a posteriori est la méthode type de correction pour tenir compte de l'inégalité des probabilités de sélection et de la non-réponse lors des enquêtes par sondage. Vue sous l'angle de la modélisation, l'application de la méthode itérative du quotient ou de la stratification a posteriori à un ensemble de covariables est étroitement liée à l'application d'un modèle de régression des réponses subordonné à ces covariables, les chiffres de population étant estimés par sommation sur la répartition connue de la population selon ces covariables. Imposer comme contraintes des covariables observées plus complètement permet d'inclure plus de renseignements pour calculer les estimations de population, mais il est bien connu que l'application de la méthode itérative du quotient à un ensemble trop grand de covariables produit des inférences dont la variabilité est inacceptable. Nous proposons d'appli-quer une méthode de stratification a posteriori à un grand sous-ensemble de variables tout en adaptant un modèle hiérarchique à la régression résultante, donc de tirer parti des points forts bien connus de l'inférence bayésienne dans le cas de modèles comptant un grand nombre de paramètres échangeables.

Tableau 2
Statistiques sommaires concernant la moyenne brute des réponses, l'estimation par la méthode itérative du quotient et les trois estimations par stratification a posteriori d'après les données agrégées des sondages. Les valeurs sommaires présentées sont la moyenne estimée des proportions de vote pour les 48 États pondérées par le nombre de personnes ayant voté dans chaque État (donc, proportion estimée des suffrages exprimés pour Bush, à l'exclusion de l'Alaska, d'Hawaii et du district de Columbia), l'erreur moyenne absolue des estimations pour les 48 États, la largeur moyenne des intervalles de 50 % pour les États et le nombre d'États pour lesquels les valeurs réelles tombent dans l'intervalle de 50 %

Résumé	Résultats réels	Moyenne non pondérée	Méthode du quotient	Effets d'État non lissés	Effets d'État nuls	Modèle hiérarchique
moyenne des suffrages exprimés	0,539	0,568	0,549	0,548	0,547	0,55
erreur absolue moyenne pour les États	–	0,056	0,066	0,049	0,048	0,035
largeur moyenne des intervalles de 50 %	–	–	–	-0,069	-0,016	-0,057
nombre d'États contenus dans l'intervalle de 50 %	–	–	–	18	3	20

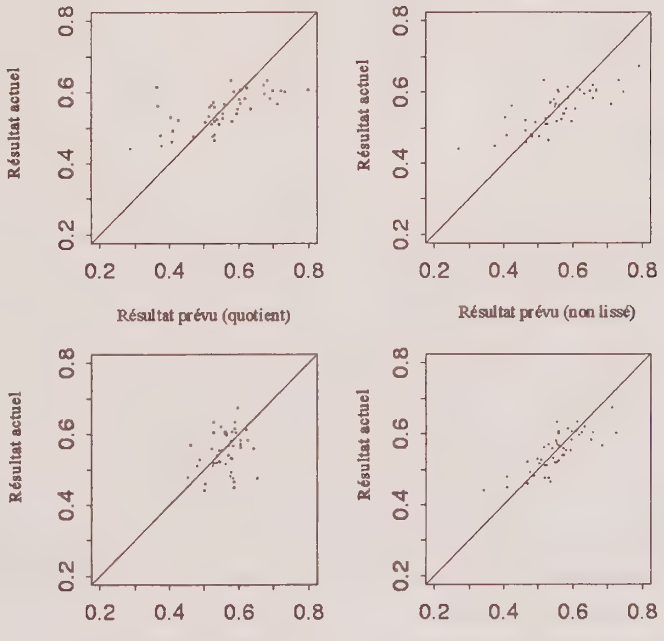


Figure 1. Résultats de l'élection selon l'État en fonction de l'estimation médiane a posteriori pour a) la méthode itérative du quotient appliqué aux variables démographiques, b) le modèle de régression incluant les indicateurs sur les États sans modèle hiérarchique, c) le modèle de régression avec les effets d'État considérés nuls et d) le modèle de régression avec adaptation d'un modèle hiérarchique aux effets d'État.

États également. Par exemple, il n'était pas réaliste de n'accorder à Bush que 46 % du soutien en Californie (durant les trois journées de sondage avant l'élection) ni 30 % seulement dans l'État de Washington. Néanmoins, à l'échelle des États-Unis, les deux estimations sont assez semblables (en fait, quand on regroupe les sept sondages, l'estimation par la méthode itérative du quotient donne des résultats un tout petit peu meilleurs), situation qui indique une fois de plus que la méthode par modélisation paraît surtout avantageuse quand on étudie des sous-ensembles de la population.

De façon étonnante, dans le cas des résultats obtenus pour l'État de Washington, l'estimation par régression fondée sur les sondages regroupés (représentées au temps «-1» sur le graphique) est plus faible que les sept estimations calculées d'après les sondages originaux. Cette observation tient au fait que les données des sondages regroupés indiquent que l'État de Washington appuie Bush moins qu'on

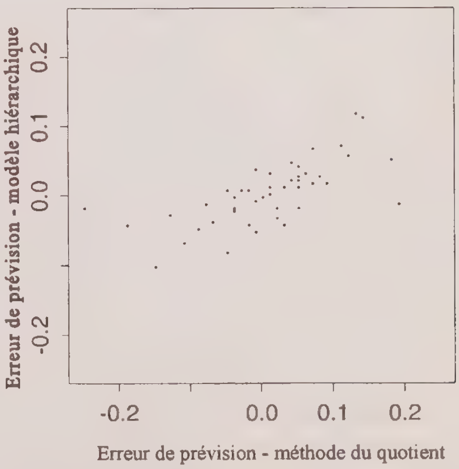


Figure 2: Diagramme de dispersion des erreurs de prévision, selon l'État, pour le modèle hiérarchique par rapport à la méthode itérative du quotient. Les erreurs produites par le modèle hiérarchique sont plus faibles par la plupart des États.

Tableau 1

Selon l'État: résultats de l'élection (proportion des votes pour les deux partis obtenue par Bush en 1988); données d'enquête (moyenne non pondérée et taille de l'échantillon) tirées des sondages regroupés; estimation par la méthode itérative du quotient en utilisant les variables de la CBS; médiane a posteriori (et intervalle interquartile, autrement dit, largeur de l'intervalle d'incertitude central de 50 %) des estimations par stratification a posteriori fondées sur les effets d'État non lissés, considérés nuls ou estimés au moyen d'un modèle hiérarchique. Les estimations sont numérotées 1, 2, 3 et 4 conformément aux descriptions de la section 3.3.

					Estimations par stratification a posteriori (et IIQ)		
État	Résultat de l'élection	Taille de l'échantillon	Moyenne non pondérée	1: Méthode du quotient itérative	2: effets d'État non lissés	3: effets d'État nuls	4: Modèle hiérarchique
AL	0,6	134	0,72	0,67	0,63 (0,05)	0,56 (0,01)	0,62 (0,05)
AR	0,57	86	0,57	0,53	0,53 (0,06)	0,60 (0,01)	0,55 (0,06)
AZ	0,61	141	0,62	0,61	0,62 (0,05)	0,56 (0,02)	0,61 (0,05)
CA	0,52	1075	0,57	0,53	0,55 (0,02)	0,53 (0,01)	0,55 (0,02)
CO	0,54	126	0,59	0,59	0,58 (0,06)	0,57 (0,01)	0,57 (0,05)
CT	0,53	103	0,53	0,55	0,52 (0,06)	0,49 (0,02)	0,51 (0,06)
DE	0,56	30	0,4	0,37	0,42 (0,11)	0,60 (0,01)	0,52 (0,08)
FL	0,61	553	0,64	0,62	0,61 (0,03)	0,62 (0,01)	0,61 (0,03)
GA	0,6	211	0,62	0,58	0,56 (0,04)	0,56 (0,01)	0,56 (0,04)
IA	0,45	102	0,38	0,38	0,38 (0,06)	0,59 (0,01)	0,41 (0,06)
ID	0,63	31	0,52	0,58	0,52 (0,12)	0,59 (0,02)	0,55 (0,08)
IL	0,51	429	0,55	0,52	0,53 (0,03)	0,52 (0,01)	0,52 (0,03)
IN	0,6	215	0,75	0,73	0,74 (0,04)	0,56 (0,01)	0,72 (0,04)
KS	0,57	105	0,72	0,71	0,71 (0,06)	0,57 (0,01)	0,68 (0,05)
KY	0,56	146	0,57	0,53	0,56 (0,05)	0,64 (0,01)	0,57 (0,05)
LA	0,55	153	0,62	0,6	0,61 (0,05)	0,54 (0,01)	0,59 (0,04)
MA	0,46	277	0,47	0,41	0,46 (0,04)	0,50 (0,02)	0,47 (0,04)
MD	0,51	207	0,52	0,5	0,49 (0,04)	0,56 (0,01)	0,50 (0,04)
ME	0,56	44	0,52	0,52	0,55 (0,10)	0,52 (0,02)	0,54 (0,08)
MI	0,54	399	0,58	0,55	0,57 (0,03)	0,54 (0,01)	0,57 (0,03)
MN	0,46	210	0,54	0,53	0,53 (0,05)	0,59 (0,01)	0,53 (0,04)
MO	0,52	235	0,46	0,43	0,46 (0,04)	0,55 (0,01)	0,47 (0,04)
MS	0,61	170	0,69	0,7	0,65 (0,04)	0,53 (0,01)	0,63 (0,04)
MT	0,53	31	0,39	0,4	0,40 (0,12)	0,58 (0,02)	0,50 (0,09)
NC	0,58	239	0,59	0,6	0,55 (0,04)	0,58 (0,01)	0,55 (0,04)
ND	0,57	54	0,56	0,56	0,55 (0,09)	0,58 (0,01)	0,56 (0,08)
NE	0,61	90	0,58	0,6	0,56 (0,07)	0,58 (0,01)	0,56 (0,06)
NH	0,63	20	0,7	0,68	0,73 (0,13)	0,53 (0,02)	0,61 (0,10)
NJ	0,57	301	0,57	0,6	0,53 (0,04)	0,46 (0,01)	0,53 (0,03)
NM	0,53	87	0,55	0,54	0,57 (0,07)	0,54 (0,02)	0,56 (0,06)
NV	0,61	19	0,68	0,8	0,67 (0,13)	0,56 (0,02)	0,60 (0,09)
NY	0,48	639	0,42	0,37	0,41 (0,03)	0,45 (0,01)	0,41 (0,02)
OH	0,55	454	0,62	0,63	0,58 (0,03)	0,55 (0,01)	0,58 (0,03)
OK	0,58	93	0,57	0,62	0,59 (0,07)	0,63 (0,01)	0,60 (0,06)
OR	0,48	111	0,5	0,47	0,50 (0,06)	0,58 (0,02)	0,52 (0,06)
PA	0,51	431	0,54	0,54	0,52 (0,03)	0,48 (0,02)	0,52 (0,03)
RI	0,44	65	0,28	0,29	0,27 (0,07)	0,50 (0,02)	0,34 (0,06)
SC	0,62	151	0,7	0,67	0,66 (0,05)	0,55 (0,01)	0,64 (0,04)
SD	0,53	52	0,54	0,51	0,53 (0,09)	0,58 (0,01)	0,54 (0,08)
TN	0,58	252	0,68	0,69	0,66 (0,04)	0,60 (0,01)	0,65 (0,03)
TX	0,56	594	0,58	0,52	0,56 (0,03)	0,60 (0,01)	0,56 (0,02)
UT	0,67	61	0,8	0,85	0,79 (0,07)	0,60 (0,02)	0,72 (0,06)
VA	0,6	255	0,69	0,72	0,67 (0,04)	0,59 (0,01)	0,66 (0,03)
VT	0,52	12	0,54	0,58	0,60 (0,19)	0,53 (0,02)	0,55 (0,11)
WA	0,49	269	0,47	0,41	0,46 (0,04)	0,58 (0,01)	0,48 (0,04)
WI	0,48	264	0,49	0,53	0,48 (0,04)	0,57 (0,01)	0,49 (0,04)
WV	0,48	79	0,48	0,52	0,48 (0,07)	0,65 (0,01)	0,53 (0,06)
WY	0,61	13	0,5	0,36	0,59 (0,17)	0,59 (0,02)	0,59 (0,10)

phiques. Les estimations au niveau de l'État produites par ce modèle devraient être meilleures que celles obtenues en appliquant la méthode itérative du quotient en regard des variables démographiques, car les estimations des π_j sont pondérées par les chiffres de population N_j plutôt que par l'effectif de l'échantillon, n_j, dans chaque État.

3. Estimation par régression en se servant uniquement des variables démographiques, et en donnant une valeur nulle aux effets d'État. En vertu de ce modèle, les réponses moyennes dans les États diffèrent uniquement à cause des variations démographiques; dans la mesure où les caractéristiques démographiques n'expliquent pas complètement la variation de l'opinion, le modèle sous-estime la variabilité d'un État à l'autre.
4. Estimation par régression en se servant des variables démographiques et en estimant les effets des 48 États au moyen d'un modèle hiérarchique (selon la notation adoptée à la section 2, $L = 4$ et K_1, K_2, K_3, K_4 = 12, 13, 12, 11). Nous nous attendons à ce que ce modèle donne les résultats les meilleurs non seulement parce que le modèle hiérarchique de régression est souple, mais aussi parce que la stratification a posteriori se fonde sur les chiffres de population N_j.

Nous ajustons chacun des modèles de régression aux données d'enquête, produisons des tirages par simulation a posteriori pour chaque coefficient (subordonnés aux $\tau_1, \tau_2, \tau_3, \tau_4$ estimés), et effectuons de nouveau la pondération d'après les données de la PUMS pour obtenir, dans chaque strate a posteriori, la proportion estimée d'électeurs enregistrés qui appuient la candidature de Bush à la présidence.

Le tableau 1 présente les estimations obtenues par la méthode itérative du quotient, les médianes et les intervalles interquartiles a posteriori pour les trois modèles, ainsi que les données sur les réponses aux sondages et les résultats réels de l'élection. Le tableau 2 donne les erreurs de prédiction au niveau national et les erreurs moyennes absolues de prédiction au niveau des États pour la méthode itérative du quotient et pour les trois modèles. Les quatre méthodes produisent pratiquement les mêmes résultats au niveau national; l'amélioration réelle des estimations grâce aux modèles se manifeste au niveau des États. La réduction de l'erreur moyenne absolue de prédiction d'environ 6 % jusqu'à 5 % peut être attribuée à l'utilisation des renseignements résultants de la stratification a posteriori, et la réduction supplémentaire jusqu'à 3,5 %, à la modélisation hiérarchique. De surcroît, les deux dernières lignes du tableau 2 montrent que les intervalles d'incertitude estimés par le modèle hiérarchique sont courts et relativement bien étalonnés (un peu moins de la moitié des valeurs vraies tombent dans les intervalles de 50 %, résultat raisonnable si l'on considère que ces intervalles tiennent compte de l'erreur d'échantillonnage, mais non des erreurs non dues à l'échantillonnage ni des variations d'opinion).

La figure 1 donne une représentation graphique, selon l'État, des résultats réels de l'élection en fonction des estimations produites par la méthode itérative du quotient et des médianes des strates a posteriori calculées pour les trois modèles. Il n'est pas étonnant de constater que le modèle hiérarchique diminue la variance, donc l'erreur d'estimation, par rétrécissement. Bien que les quatre méthodes permettent de corriger de pratiquement la même grandeur le biais qui entache l'estimation au niveau national, elles ont des effets différents au niveau de l'État, le modèle hiérarchique étant celui qui produit les résultats les meilleurs. La figure 2 permet de comparer les erreurs de prédiction résultant de l'application du modèle hiérarchique et de la méthode itérative du quotient pour produire les estimations pour les États.

Fait intéressant, le modèle hiérarchique ne semble pas rapprocher suffisamment les données de la moyenne nationale puisque, comme le montre la figure d, le résultat actuel de l'élection est plus élevé que prévu pour les valeurs que l'on prévoyait faibles et plus faibles que prévu pour celles que l'on prévoyait élevées. Le *sous*-rétrécissement signifie que la valeur des paramètres estimés $\hat{\tau}_l$ est probablement *plus grande* que leur valeur réelle, situation qui pourrait être due à une courbe de non-réponse non négligeable, variant d'un État à l'autre, si bien que la variabilité observée des proportions au niveau de l'État résulte de la variation de la courbe de non-réponse en plus de la variation réelle de la moyenne des opinions (consulter Little et Gelman 1996, pour un examen de cet exemple, ainsi que Krieger et Pfeffermann 1992, pour un traitement plus général). On pourrait quantifier le sous-rétrécissement en comparant le niveau de rétrécissement estimé au niveau jugé optimal, mais ceci n'est faisable qu'après avoir observé les valeurs réelles.

On peut aussi comparer les modèles en les ajustant individuellement à chaque sondage et en examinant la stabilité des estimations au cours d'une période brève. Il s'agirait-là d'un moyen raisonnable d'étudier les modèles dans la situation, courante, où on ne connaît jamais les moyennes réelles de population. La figure 3 montre, pour chacun des sept sondages, les estimations obtenues par la méthode itérative du quotient et grâce au modèle hiérarchique. (Au moment de la modélisation individuelle des enquêtes, nous avons utilisé une variance hiérarchique commune pour les 48 États, car nous ne disposions pas de données suffisantes pour obtenir des estimations fiables du maximum de vraisemblance pour les quatre régions séparément d'après les données de chaque sondage.) Les résultats sont présentés pour l'ensemble des États-Unis et pour trois États représentatifs, à savoir la Californie (grand État), l'État de Washington (État de taille moyenne) et le Nevada (petit État). Par souci de commodité, la représentation graphique montre aussi les estimations calculées d'après les données agrégées des sept sondages et les résultats réels de l'élection. Pour chacun des États, l'estimation grâce au modèle hiérarchique varie moins que celle obtenue par la méthode itérative du quotient. C'est pour le Nevada, où l'effectif des échantillons des divers sondages était si faible que les estimations par la méthode itérative du quotient se réduisaient à 0 ou à 1 dans la plupart des cas, que la tendance est la plus nette, mais la supériorité du modèle hiérarchique est manifeste dans le cas des autres

Alaska, seuls les 48 États contigus figurent dans le modèle. Bien qu'il soit inclus dans les enquêtes, l'État de Washington, D.C. est exclu de l'analyse. En effet, les préférences en matière de vote y diffèrent tellement de celles observées dans les autres États qu'un modèle linéaire généralisé adapté aux 48 États ne serait pas aussi bien adapté à cet État et, les données qu'on y collecterait influeraient donc indûment sur les résultats obtenus pour les états. Puisqu'on dispose de moins d'observations pour les petits états et que la variation du soutien estimé pour Bush d'un sondage à l'autre est similaire à la variabilité d'échantillonnage binomial (telle que mesurée par le test χ^2 de l'égalité des proportions d'électeurs qui appuient Bush dans les sept sondages), nous regroupons les données de tous les sondages.

La CBS détermine les coefficients de pondération d'enquête par application de la méthode itérative du quotient aux variables suivantes, avec les classifications implicites pour la non-réponse à une question indiquée entre crochets:

région de recensement:	nord-est, sud, centre nord, ouest.
sexe:	masculin, féminin.
groupe ethnique:	noir, [blanc/autre].
âge:	de 18 à 29 ans, de 30 à 44 ans, [de 45 à 64 ans], 65 ans et plus.
niveau de scolarité:	pas de diplôme d'études secondaires, [diplôme d'études secondaires], certaines études collégiales, diplôme collégial.

L'application de la méthode itérative du quotient englobe tous les effets importants plus les interactions sexe × groupe ethnique et âge × éducation. Nous incluons toutes ces variables à titre d'effets constants dans le modèle de régression logistique, et nous excluons de l'analyse les répondants, assez rares, qui ne produisent pas de réponse pour une variable démographique quelconque. Les coefficients de pondération calculés par la CBS tiennent aussi compte des nombres de lignes téléphoniques et d'adultes dans le ménage, car ils ont une incidence sur les probabilités d'échantillonnage; cependant, ces éléments n'ont qu'un effet mineur sur les estimations de la préférence pour l'un ou l'autre candidat à la présidence (consulter Little 1996, chapitre 3) et nous ne les incluons pas dans le modèle. Le lecteur trouvera d'autres renseignements sur les méthodes d'enquête et de correction appliquées par la CBS dans Voss, Gelman et King (1995).

Notre modèle va au-delà de l'analyse effectuée par la CBS, car il comprend des indicateurs des effets aléatoires liés aux 48 États, regroupés en quatre lots correspondant aux quatre régions de recensement. Nous vérifions la performance du modèle en comparant les estimations obtenues pour chaque État aux résultats réels de l'élection présidentielle. (Les sondages d'opinion effectués juste avant l'élection sont des indicateurs fiables du résultat réel de l'élection; consulter, p. ex., Gelman et King 1993.) Nous comparons aussi la stabilité des estimations fondées sur les résultats de divers sondages au cours d'une brève période.

3.2 Chiffres de population pour la stratification a posteriori

Afin de faire une stratification a posteriori en regard de toutes les variables susmentionnées, ainsi que de l'État, nous devons connaître la répartition agrégée de population pour les variables démographiques dans chaque État, c'est-à-dire les totaux de population N_j pour chacune des $2 \times 2 \times 4 \times 48$ cellules définies par sexe × groupe ethnique × âge × État. Puisque les électeurs enregistrés sont la population cible, nous devrions nous fonder sur la répartition de cette population. En tant qu'approximation, nous utilisons les totalisations croisées provenant des données de la *Public Use Micro Survey* (PUMS) pour tous les citoyens de 18 ans et plus. Les données de la PUMS contiennent des enregistrements pour 5 % des unités de logement aux États-Unis et pour les personnes qui les habitent, soit plus de 12 millions de personnes et plus de 5 millions d'unités de logement. Ces données produisent un échantillon stratifié des 15,9 % d'unités de logement environ qui ont reçu un questionnaire détaillé à l'occasion du Recensement de 1990. Les personnes qui vivent en établissements ou dans d'autres logements collectifs sont également incluses dans l'échantillon. Les poids sont calculés, tant pour l'unité de logement que pour les personnes qui l'occupent, d'après les probabilités d'échantillonnage et les corrections apportées aux totaux du recensement pour les variables incluses dans le questionnaire abrégé. Nous utilisons les données pondérées de la PUMS pour estimer N_j pour chaque strate a posteriori et nous ne tenons pas compte de l'erreur d'échantillonnage dans ces chiffres. Les chiffres pondérés tirés de la PUMS sont fort semblables à ceux provenant de la stratification a posteriori auxquels la CBS a appliqué la méthode itérative du quotient (voir Little 1996, chapitre 3).

3.3 Résultats

Nous présentons les résultats pour quatre méthodes appliquées aux données agrégées de sept sondages:

1. Estimation classique par la méthode itérative du quotient selon les variables démographiques (région, sexe, groupe ethnique, âge, niveau de scolarité, sexe × groupe ethnique et âge × niveau de scolarité). Cette méthode est fort semblable à la méthode de pondération utilisée par la CBS. Pour l'estimation des résultats selon l'État, nous calculons les moyennes pondérées pour chaque État, d'après les poids obtenus par la méthode itérative du quotient.
2. Estimation par régression en se servant des variables démographiques ainsi que des indicateurs sur les États, sans modèle hiérarchique (c.-à-d. régression en supposant les effets d'État constants). Cette méthode est fort semblable à l'ajustement itératif proportionnel couvrant les États ainsi que les variables démogra-

– Le fait de n'inclure aucune variable explicative dans le modèle (autrement dit à poser que X est simplement un vecteur de 1) mène à l'estimation de la moyenne d'échantillon $\bar{y}$.

Pour une discussion plus détaillée de la relation entre les estimations par pondération et la stratification a posteriori, consulter Holt et Smith (1979), ainsi que Little (1993).

2.3 Modèle de régression hiérarchique pour regroupement partiel

Si le nombre de cellules est grand, aucune option susmentionnée ne permet d'utiliser efficacement les renseignements fournis par les catégories (par exemple, la stratification a posteriori simple produit des estimations qui sont trop variables; toutefois, si nous excluons les variables explicatives pour un grand nombre de catégories, nous éliminons des renseignements importants). Pour remédier à cette situation, nous effectuons un groupement partiel des cellules en ajustant un modèle à effets mixtes (consulter, par exemple, Clayton 1996). Nous représentons le vecteur β par $(\alpha, \gamma_1, ..., \gamma_L)$, où α est un sous-vecteur de coefficients non groupés et où chaque γ_l, pour $l = 1, ..., L$, est un sous-vecteur de coefficients (γ_{kl}) auxquels nous ajustons un modèle hiérarchique:

$$\gamma_{kl} \overset{\text{ind}}{\sim} N(0, \tau_l^2), k = 1, ..., K_l.$$

Donner à τ_l une valeur nulle (0) équivaut à exclure un ensemble de variables; donner à τ_l une valeur infinie (∞) équivaut à une distribution a priori non informative en regard des paramètres γ_{kl}.

Étant donné les réponses y_i dans les catégories j, nous construisons une matrice C de catégorisation $n \times J$ pour laquelle $C_{ij} = 1$ si le répondant i se trouve dans la cellule j. Posons que $Z = CX$. On peut alors écrire l'équation du modèle (1) sous la forme d'un modèle hiérarchique de régression logistique de la façon suivante:

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = Z\beta$$

$$\beta \sim N(0, \textstyle\sum_\beta),$$

où $\sum_\beta^{-1}$ est une matrice diagonale dont tous les éléments de α sont nuls, suivis de τ_l^{-2} pour chaque élément de γ_l, pour chaque l. Nous représentons par p_i la probabilité correspondant à l'unité i, de façon à la distinguer de π_j, c'est-à-dire la probabilité agrégée correspondant à la catégorie j. Consulter Nordberg (1989), ainsi que Belin, Diffendal, Mack, Rubin, Schafer et Zaslavsky (1993) pour une discussion générale des modèles hiérarchiques de régression logistique applicables aux données d'enquête.

2.4 Inférence en vertu du modèle

Pour faire des inférences au sujet des paramètres à l'échelle de la population, nous adoptons la stratégie empirique de Bayes, c'est-à-dire premièrement, estimer les hyperparamètres τ_l, étant donné les valeurs de y; deuxièmement, faire une inférence bayésienne en ce qui concerne les coefficients de régression β, étant donné y et les τ_l estimés; troisièmement, calculer les inférences pour le vecteur des moyennes des cellules $\pi = \text{logit}^{-1}(X\beta)$; quatrièmement, calculer les inférences pour les paramètres à l'échelle de la population en additionnant les $N_j\pi_j$. Nous considérons cette méthode comme une approximation de l'analyse bayésienne complète, qui consiste à faire la moyenne sur les paramètres τ_l. Les deux méthodes diffèrent surtout quand on estime les composantes τ_l de façon imprécise ou qu'on ne peut les distinguer de 0 (consulter, par exemple, Gelman et coll. 1995, section 5.5). Dans l'exemple examiné ici, le problème ne se pose pas, car l'estimation des diverses composantes montre clairement que ces dernières diffèrent de 0. Si cela n'était pas le cas, cela vaudrait sûrement la peine de pousser plus loin l'effort de programmation afin d'effectuer une analyse bayésienne complète. Toutefois, la présente étude vise à examiner l'efficacité de la combinaison de la modélisation hiérarchique à la stratification a posteriori, plutôt que les différences techniques assez mineures entre les analyses bayésiennes empiriques et non empiriques.

Le rétrécissement des estimations dans les cellules a lieu à la deuxième étape et son importance dépend de la taille de l'échantillon n_j et des valeurs de $\bar{y}_j$. Le rétrécissement est d'autant plus important que les valeurs de n_j sont faibles et que les valeurs de $\bar{y}_j$ s'écartent des prédictions fondées sur le modèle de régression logistique. En outre, le rétrécissement est plus important si les paramètres τ_l sont petits. Ainsi, un lot de coefficients γ_l dont le pouvoir prédictif est faible sera réduit de façon à tendre vers zéro dans l'estimation, parce qu'il sera estimé que τ_l a une valeur faible. Cette méthode permet d'inclure un grand nombre de coefficients dans le modèle hiérarchique sans augmenter trop la variabilité des estimations des grandeurs à l'échelle de la population.

3. APPLICATION: VENTILATION DES DONNÉES D'ENQUÊTES NATIONALES SELON L'ÉTAT

3.1 Données d'enquête

Nous appliquons la méthodologie susmentionnée pour déterminer, au niveau de l'État, les résultats de sept sondages d'opinion nationaux effectués par le réseau de télévision CBS auprès des électeurs enregistrés durant les deux semaines précédant directement l'élection présidentielle de 1988 aux États-Unis. Conformément à la notation générale que nous avons adoptée, nous assignons $y_i = 1$ aux partisans de Bush et $y_i = 0$ aux partisans de Dukakis; nous éliminons les enquêtés qui n'ont exprimé aucune opinion (environ 15 % du total; conformément à la pratique courante, nous comptons les répondants qui «penchent» vers un candidat comme des partisans à part entière). Puisqu'aucune donnée n'a été collectée à Hawaii ni en

Dans le présent article, nous décrivons un modèle hiérarchique de régression logistique conçu pour estimer une variable binaire par stratification a posteriori. Comparativement à la stratification a posteriori type, ce modèle permet d'utiliser un nombre nettement plus grand de catégories, donc des renseignements beaucoup plus détaillés sur la population. En pratique, la méthode est surtout avantageuse dans le cas des petits sous-groupes de population. Nous l'appliquons aux résultats, au niveau de l'État, d'un ensemble de sondages d'opinion préélectoraux effectués aux États-Unis. Le choix de cet exemple nous permet notamment de vérifier nos inférences au moyen d'une source externe, en les comparant aux résultats électoraux au niveau de l'État. En annexe, nous décrivons le calcul du modèle hiérarchique au moyen d'un algorithme EM d'espérance approximative et de maximisation.

2. MODÈLE

2.1 Renseignements sur l'échantillonnage et la stratification a posteriori

Considérons une subdivision de la population en R variables nominales, où la r-ième variable possède J_r niveaux, ce qui donne un total de $J = \Pi_{r=1}^{R} J_r$ catégories (cellules), que nous annotons $j = 1, ..., J$. Supposons qu'on connaît N_j, c'est-à-dire le nombre d'unités de population dans la catégorie j, pour toutes les valeurs de j. Représentons par y une réponse binaire que l'on veut étudier et représentons par π_j, la réponse moyenne de la population dans chaque catégorie j. Alors, la moyenne globale de la population est $\bar{Y} = \sum_j N_j \pi_j / \sum_j N_j$. Supposons que la population est suffisamment grande pour qu'on puisse ignorer toutes les corrections ayant trait aux populations finies.

Effectuons maintenant une enquête par sondage en vue d'estimer $\bar{Y}$ (et peut-être certains autres regroupements des π_j). Pour chaque j, représentons par n_j le nombre d'unités dans la catégorie j de l'échantillon. En la subordonnant aux variables explicatives R, émettons l'hypothèse qu'on peut ignorer l'impact de la non-réponse (Rubin, 1976). Donc, les variables R devraient inclure tous les renseignements nécessaires pour calculer les poids d'enquête, ainsi que toute autre variable susceptible de fournir des renseignements sur y.

Dans le cas de l'exemple exposé à la section 3, nous catégorisons la population d'adultes dans les 48 États américains contigus d'après $R = 5$ variables, à savoir l'état de résidence, le sexe, le groupe ethnique, l'âge et le niveau de scolarité, avec $(J_1, ..., J_5) = (48, 2, 2, 4, 4)$. (Les variables du groupe ethnique, de l'âge et du niveau de scolarité sont discrétisées chacune en 4 catégories, comme on le décrit à la section 3.1.) Les $J = 3\ 072$ catégories varient de «Alabama, homme, noir, de 18 à 29 ans, sans diplôme d'études secondaires» à «Wyoming, femme, non noire, 65 ans et plus, diplôme collégial». D'après les données du Recensement des États-Unis, nous pouvons calculer de bonnes estimations de N_j dans chacune de ces catégories.

Nous considérerons des estimations pour l'ensemble de la population (obtenues en calculant la somme pour les 3 072 catégories) ainsi que des estimations par État (en calculant séparément la somme de 64 catégories dans chaque État). Puisque, dans le cas d'un échantillon d'enquête de taille raisonnable, il est impossible d'obtenir des estimations indépendantes des réponses moyennes π_j pour des catégories j distinctes (en fait, la plupart des catégories sont vides ou ne contiennent qu'un seul répondant), nous devons modéliser les π_j pour pouvoir faire la stratification a posteriori, et donc nous servir des effectifs connus des catégories N_j. La stratification a posteriori offre l'avantage (éventuel) d'apporter une correction pour la variation du taux de non-réponse d'une catégorie à l'autre.

2.2 Modèles de régression dans le contexte de la stratification a posteriori

On peut créer un modèle de régression logistique pour déterminer la probabilité π_j que les répondants de la catégorie j disent «oui». Il aura la forme

$$\text{logit}(\pi_j) = X_j \beta, \tag{1}$$

où X est une matrice de variables explicatives et où X_j représente la j-ième rangée de X. Si nous supposons que la distribution a priori est uniforme en regard de β, alors, en vertu du modèle susmentionné, l'inférence bayésienne pour différents choix de X correspond de près à divers schémas classiques de pondération. Ces correspondances, que nous présentons ci-après, sont générales et s'appuient sur la linéarité du modèle supposé (autrement dit, $X_j \beta$ dans (1)). (Dans le cas des données binaires étudiées dans le présent article, les estimations classiques et bayésiennes avec une distribution a priori uniforme ne sont pas identiques étant donné la transformation logistique non linéaire représentée par (1), mais, pour de grands échantillons, les écarts sont minimes.)

Les modèles qui suivent correspondent aux estimations classiques par stratification a posteriori les plus courantes.

- Faire correspondre X à la matrice d'identité $J \times J$ équivaut à pondérer chaque unité de la cellule j par N_j / n_j, autrement dit, à effectuer une stratification a posteriori simple. Il est bien connu que cette méthode ne donne de bons résultats que si les n_j sont suffisamment grands (et ne marche pas du tout si $n_j = 0$ pour certains j).
- Si nous faisons correspondre X à la matrice de variables explicatives $J \times (\sum_{r=1}^{R} J_r)$ pour chaque variable, alors, l'estimation de $\bar{Y}$ correspond à peu près à celle obtenue par application de la méthode itérative du quotient entre les marges unidimensionnelles pour toutes les R.
- Inclure diverses interactions dans X revient à inclure ces interactions dans l'ajustement proportionnel itératif. De façon plus générale, supposer que X présente une «structure» quelconque équivaut à regrouper d'une certaine façon les strates a posteriori.

Techniques d'enquête, décembre 1997
Volume 23, n° 2, pages 135-145
Statistique Canada

Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique

ANDREW GELMAN et THOMAS C. LITTLE[1]

RÉSUMÉ

La stratification a posteriori est une méthode appliquée couramment pour tenir compte de l'inégalité des probabilités d'échantillonnage et de la non-réponse lors des enquêtes par sondage. Cette méthode consiste à subdiviser la population en plusieurs catégories, à estimer la répartition des réponses dans chaque catégorie, puis, à donner à chaque catégorie un poids proportionnel à sa taille dans la population. Nous considérons la stratification a posteriori comme un cadre de référence général englobant de nombreux scénarios de pondération utilisés dans le domaine de l'analyse d'enquête (consulter Little 1993). Nous construisons un modèle de régression logistique hiérarchique pour déterminer la moyenne conditionnelle d'une variable de réponse binaire subordonnée à des cellules, ou catégories, de stratification a posteriori. Le modèle hiérarchique permet d'inclure un nombre beaucoup plus grand de cellules que les méthodes classiques, donc, d'introduire beaucoup plus de renseignements sur la population, tout en incluant tous les renseignements qui sous-tendent l'inférence lors de l'échantillonnage d'enquête. Donc, nous combinons la méthode de modélisation appliquée fréquemment à l'estimation des petites régions aux renseignements sur la population utilisés à l'étape de la stratification a posteriori. Nous appliquons la méthode à un ensemble de sondages d'opinion préélectoraux effectués aux États-Unis, dont les données sont stratifiées a posteriori selon l'État et selon les variables démographiques habituelles. Nous évaluons les modèles graphiquement en comparant les résultats qu'ils produisent à ceux des élections au niveau de l'État.

MOTS CLÉS: Inférence Bayésienne; prévision électorale; non-réponse; sondages d'opinion; enquêtes par sondage.

1. INTRODUCTION

Le fait de fonder entièrement ou principalement la pondération sur la stratification a posteriori, expression qui désigne généralement toute méthode d'estimation visant à rajuster les chiffres d'après les totaux calculés pour l'ensemble de la population, est une pratique courante dans le cas des sondages d'opinion. Essentiellement, la méthode se résume à répartir la population en un certain nombre de catégories à l'intérieur desquelles les résultats de l'enquête sont analysés comme s'ils étaient obtenus selon un plan d'échantillonnage aléatoire simple. L'étape de stratification a posteriori consiste à estimer les paramètres à l'échelle de la population en faisant la moyenne des estimations dans les catégories, après avoir donné à celles-ci un poids proportionnel à leur taille relative dans la population. Ordinairement, on définit les catégories de la stratification a posteriori d'après les caractéristiques démographiques (sexe, âge, *etc.*) et d'après toute variable utilisée dans la stratification. Un autre niveau de complication, que nous n'abordons pas ici, surviendrait en cas d'échantillonnage en grappes.

La définition des catégories de la stratification a posteriori pose une difficulté fondamentale. Il est souhaitable de diviser la population en un grand nombre de petites catégories, afin que l'hypothèse selon laquelle l'échantillonnage est aléatoire simple dans chaque catégorie soit raisonnable. Toutefois, si le nombre de répondants par catégorie est faible, il est difficile d'estimer avec précision la réponse moyenne dans chaque catégorie. Par exemple, si on stratifie a posteriori selon le sexe, le groupe ethnique, l'âge, le niveau de scolarité et la région des États-Unis, certaines cellules de l'échantillon pourraient être vides, tandis que d'autres ne contiendraient qu'un ou deux répondants.

Un moyen général de résoudre ce problème consiste à modéliser les réponses en subordonnant le modèle aux variables de stratification a posteriori (consulter Little 1993). Par exemple, pour corriger les données en fonction de plusieurs variables démographiques, on applique généralement la méthode itérative du quotient entre des marges unidimensionnelles ou bidimensionnelles (c.-à-d. un ajustement proportionnel itératif, Deming et Stephan 1940). Cet exercice correspond essentiellement à faire une stratification a posteriori couvrant entièrement le tableau multidimensionnel, mais en se servant d'un modèle des réponses, subordonné aux variables démographiques, qui donne une valeur nulle aux interactions de niveau supérieur. Les méthodes fondées sur les poids de lissage peuvent aussi être considérées comme des stratifications a posteriori, avec modèles de réponses correspondants (consulter Little 1991). Quand les catégories de la stratification a posteriori sont conformes à une structure hiérarchique (par exemple, personnes dans un État aux États-Unis), on peut améliorer l'efficacité de l'estimation en ajustant un modèle hiérarchique (p. ex., Lazzeroni et Little 1997). Dans le contexte connexe de l'estimation par régression, Longford (1996) montre que les modèles hiérarchiques linéaires permettent d'améliorer la précision des estimations des petites régions fondées sur des données d'enquête par sondage.

[1] Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027 et Thomas C. Little, Morgan Stanley Dean Witter, New York, NY.

petites régions en s'appuyant sur les variables de résultats ordinales lorsqu'on a des raisons de craindre que l'utilisation d'un modèle ordinal n'aboutisse à des estimations négatives pour certaines de ces probabilités.

REMERCIEMENTS

Cette étude a bénéficié de l'aide financière du CRSNG du Canada. L'auteur remercie le rédacteur associé et les lecteurs pour leurs commentaires et leurs suggestions utiles.

BIBLIOGRAPHIE

ALBERT, J.H., et CHIB, S. (1993). Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association*, 88, 669-679.

ANDERSON, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, 46, 1-30.

BETHLEHEM, J.G., KELLER, W.J., et PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

BRESLOW, N.E., et CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

BRESLOW, N.E., et LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82, 81-91.

CAMPBELL, M.K., et DONNER, A. (1989). Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association*, 84, 587-591.

CAMPBELL, M.K., DONNER, A., et WEBSTER, K.M. (1991). Are ordinal models useful for classification? *Statistics in Medicine*, 10, 383-394.

CARLIN, B.P., et GELFAND, A.E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85, 105-114.

CRESSIE, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquête*, 18, 83-103.

CROUCHLEY, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, 90, 489-498.

DEMPSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

DEMPSTER, A.P., et TOMBERLIN, T.J. (1980). The analysis of census undercount from a postenumeration survey. *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.

FARRELL, P.J. (1991). Empirical Bayes Estimation of Small Area Proportions. Thèse de doctorat, Department of Management Science, McGill University, Montréal, Québec, Canada.

FARRELL, P.J., MacGIBBON, B., et TOMBERLIN, T.J. (1997a). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

FARRELL, P.J., MacGIBBON, B., et TOMBERLIN, T.J. (1997b). Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

GONZALES, M.E. (1973). Use and evaluation of synthetic estimation. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons Inc.

LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 581-590.

LAIRD, N.M., et LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82, 739-750.

MacGIBBON, B., et TOMBERLIN, T.J. (1989). Estimation de proportions pour petites régions par des méthodes empiriques de Bayes. *Techniques d'enquête*, 15, 247-262.

MALEC, D., SEDRANSK, J., et TOMPKINS, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. In *Case Studies in Bayesian Statistics*, (Éds. C. Gatsonis, J.S. Hodges, R. Kasf, et N.D. Singpurwalla). New York: Springer Verlag.

McCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109-142.

PRASAD, N.G.N., et RAO, J.N.K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, 85, 163-171.

RIPLEY, B.D., et KIRKLAND, M.D. (1990). Iterative simulation methods. *Journal of Computational and Applied Mathematics*, 31, 165-172.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 74, 1-12.

STROUD, T.W.F. (1991). Hierarchical Bayes predictive means and variances with application to sample survey inference. *Communications in Statistics, Theory and Methods*, 20, 13-36.

TOMBERLIN, T.J. (1988). Predicting accident frequencies for drivers classified by two factors. *Journal of the American Statistical Association*, 83, 309-321.

UNITED STATES BUREAU OF THE CENSUS (1984). Census of the Population, 1950: Public Use Microdata Sample Technical Documentation, édité par J.G. Keane, Washington, D.C.

WONG, G.Y., et MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

ZEGER, S.L., et KARIM, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79-86.

Pour chaque catégorie de revenu, le biais relatif moyen et le biais relatif absolu moyen de la racine carrée de $\widehat{\mathrm{Var}}(\hat{p}_{im+})$, utilisés à titre d'estimation de la valeur empirique de la REQM, sont présentés au tableau 1 pour les modèles multinomial et ordinal. Le biais relatif moyen correspond simplement à la moyenne, pour l'ensemble des régions locales échantillonnées, des valeurs obtenues lorsque la différence découlant de la soustraction de la valeur empirique de la REQM pour la i-ième région locale de la valeur moyenne de la racine carrée de $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ pour cette région, répétée pour l'ensemble des 500 échantillons de simulation, est divisée par la valeur empirique de la REQM. Le biais absolu moyen est défini d'une façon analogue, mais en utilisant cette fois la valeur absolue de chaque différence. Le tableau présente également les moyennes semblables correspondant aux mesures ajustées par la méthode bootstrap de la variabilité $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$. Pour les modèles logistiques multinomial et ordinal, le biais relatif moyen et le biais relatif absolu moyen des estimations de la variabilité ajustées par la méthode bootstrap sont sensiblement plus faibles que leurs contreparties obtenues par la méthode naïve pour l'ensemble des trois catégories de revenu. En outre, ces statistiques sommaires de la moyenne ajustée par la méthode bootstrap sont toutes très petites, ce qui porte à conclure que les estimations de la variabilité ajustée par la méthode bootstrap peuvent prendre en compte la majeure partie de l'incertitude qui découle de l'utilisation d'une estimation de la distribution des effets aléatoires.

Pour chaque région locale échantillonnée, les taux de couverture naïfs et ajustés par la méthode bootstrap, fondés sur des estimations par intervalles de confiance à 95 %, ont été calculés pour plus de 500 échantillons avec chacun des deux modèles et chacune des trois catégories de revenu. Pour l'ensemble des combinaisons de catégorie de revenu et de modèle, les taux de couverture ajustés par la méthode bootstrap pour les régions locales individuelles variaient de 92,2 à 97,6 %. Puisque la borne approximative pour l'erreur de Monte Carlo est 3 $\sqrt{(0{,}96)(0{,}05)/500}$, ou 0,029, tous les taux de couverture ajustés par la méthode bootstrap se trouvent en-deçà de 3 écarts-types de 95 %.

Pour chaque combinaison de modèle et de catégorie de revenu, on a calculé la moyenne des taux de couverture appropriés sur l'ensemble des régions locales échantillonnées pour obtenir les taux moyens naïfs et ajustés par la méthode bootstrap présentés au tableau 1. On peut tirer de ces résultats un certain nombre d'observations valables pour chacune des catégories de revenu. Pour les modèles multinomial et ordinal, les taux de couverture moyens pour les intervalles ajustés par la méthode bootstrap sont beaucoup plus près du taux nominal de 95 % que ceux associés aux intervalles naïfs. Toutefois, les taux de couverture moyens naïfs et ajustés par la méthode bootstrap pour le modèle ordinal sont légèrement meilleurs que leurs contreparties du modèle multinomial. C'est ce que l'on observe également pour l'écart absolu moyen des deux catégories de taux de couverture par rapport au taux nominal de 95 %. L'écart absolu moyen des taux de couverture naïfs par rapport au taux nominal de 95 % correspond simplement à la moyenne, pour l'ensemble des régions locales échantillonnées, des valeurs absolues de la différence obtenue lorsqu'on soustrait le taux nominal de 95 % des taux de couverture naïfs pour les régions locales échantillonnées sur l'ensemble des 500 échantillons de simulation. L'écart absolu moyen des taux de couverture ajustés par la méthode bootstrap par rapport au taux nominal de 95 % est défini d'une manière analogue.

Vingt-deux des régions locales n'ont pas été échantillonnées. On a également obtenu pour ces régions des estimations de la proportion des sujets appartenant à chacune des catégories de revenu à l'aide des modèles multinomial et ordinal. Les résultats obtenus étaient semblables à ceux des régions locales échantillonnées. Toutefois, la performance des modèles s'est détériorée quelque peu puisque les régions locales non échantillonnées constituent un échantillon restant. Pour une évaluation détaillée des résultats correspondant aux régions locales non échantillonnées, voir Farrell et coll. (1997a).

On a également comparé les estimations pour les 3 catégories de revenu fondées sur les micro-données, $\hat{p}_{im+}$, à celles fondées sur les statistiques sommaires des régions locales, $\tilde{p}_{im+}$, pour chacun des modèles. Pour les deux modèles, les résultats obtenus pour $\tilde{p}_{im+}$ étaient heureusement proches de ceux obtenus à l'aide de $\hat{p}_{im+}$, même si ceux obtenus pour $\hat{p}_{im+}$ étaient légèrement meilleurs. Farrell et coll. (1997b) ont obtenu des résultats semblables en procédant à une comparaison détaillée de $\hat{p}_{im+}$ et de $\tilde{p}_{im+}$ pour une variable de résultats binômiale.

4. CONCLUSION

En utilisant des modèles logistiques multinomial et ordinal, on a adapté la démarche empirique de Bayes proposée par Farrell et coll. (1997a) à l'estimation des proportions de petites régions à partir des données de résultats binômiales afin de prendre en compte les variables appartenant à plus de deux catégories. On a ainsi déterminé que la performance de la méthode est maintenue pour les données de résultats appartenant à des catégories multiples.

Pour comparer les estimations des proportions pour petites régions fondées sur une variable ordinale à l'aide des modèles logistiques multinomial et ordinal, on a appliqué les méthodes empiriques de Bayes fondées sur ces deux modèles à des données issues du recensement américain de 1950 en cherchant à prédire, pour une petite région, la proportion des personnes appartenant à diverses catégories d'une variable de réponses ordinale représentant le niveau de revenu. Les estimations fondées sur le modèle ordinal ne sont que légèrement meilleures en ce qui a trait au biais du plan de sondage, à la REQM empirique et aux taux de couverture. En outre, le modèle logistique ordinal se distingue particulièrement par le fait que la contrainte $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ doit être respectée pour que $\pi_{ij(m+1)} \geq 0$. Puisque les résultats des modèles multinomial et ordinal sont très semblables, on pourrait utiliser un modèle multinomial pour l'estimation des proportions de

après 150 échantillons supplémentaires. Le tableau 1 comprend les statistiques sommaires obtenues pour les 200 premiers échantillons (entre parenthèses) pour fins de comparaison.

Pour chacune des catégories de revenu, deux statistiques sommaires présentées au tableau 1 ont été évaluées pour comparer le biais dû au plan de sondage de $\hat{p}_{im+}$ pour les modèles multinomial et ordinal, le biais moyen de $\hat{p}_{im+}$ et le biais absolu moyen de $\hat{p}_{im+}$. Le biais moyen correspond simplement à la moyenne pour l'ensemble des régions locales échantillonnées des différences obtenues lorsque la proportion réelle, p_{im+}, pour la i-ième région locale est soustraite de l'estimation ponctuelle moyenne pour la région sur les 500 échantillons de simulation. Le biais absolu moyen est défini d'une façon similaire, mais en utilisant cette fois la valeur absolue de chaque différence. D'une manière générale, les résultats obtenus pour ces deux statistiques sommaires étaient légèrement meilleurs dans le cas du modèle ordinal, sans égard à la catégorie de revenu examinée. Toutefois, le modèle multinomial a laissé voir un biais moyen quelque peu plus faible pour $\hat{p}_{im+}$, pour la catégorie des personnes à faible revenu.

Pour chaque région locale échantillonnée, les valeurs empiriques la racine carrer de l'erreur quadratique moyenne (REQM) ont été calculées pour les 500 échantillons de simulation avec chacun des deux modèles et chacune des trois catégories de revenu. Pour chaque combinaison de modèle et de niveau de revenu, les valeurs empiriques appropriées de la REQM ont été calculées pour l'ensemble des régions locales échantillonnées, pour donner les valeurs empiriques moyennes de la REQM présentées au tableau 1. Ici encore, le modèle ordinal donne une performance légèrement meilleure pour l'ensemble des trois catégories de revenu.

Pour examiner la réduction de la valeur empirique de la REQM lorsqu'on utilise une estimation fondée sur la modélisation au lieu d'une méthode classique de conception non biaisée, on a calculé les valeurs empiriques moyennes de la REQM analogues à celles du tableau 1 fondées sur les 500 échantillons, en utilisant les proportions observées des échantillons de régions locales au lieu de $\hat{p}_{im+}$. Les valeurs empiriques moyennes de la REQM obtenues étaient sensiblement plus grandes (0,0617, 0,0564 et 0,0311 pour les catégories à revenu faible, moyen et élevé respectivement) que celles fondées sur $\hat{p}_{im+}$ et ce, pour les deux modèles.

Le tableau 1 comprend également des statistiques sommaires portant sur l'ensemble des régions locales échantillonnées et qui mettent en rapport les mesure naïve et celles obtenues par la méthode bootstrap de la variabilité de $\hat{p}_{im+}$, ainsi que la valeur empirique moyenne de la REQM.

Tableau 1

Statistiques sommaires moyennes fondées sur 500 échantillons de simulation pour les modèles logistique multinomial et ordinal, sur l'ensemble des petites régions échantillonnées pour chacune des catégories de revenu. Les statistiques sommaires moyennes obtenues pour les 200 premiers échantillons de simulation sont indiquées entre parenthèses, pour fins de comparaison

Moyenne	Faible revenu		Revenu moyen		Revenu élevé	
	Multinomial	Ordinal	Multinomial	Ordinal	Multinomial	Ordinal
Biais de $\hat{p}_{im+}$	-0,0004 (-0,0004)	-0,0005 (-0,0006)	-0,0007 (-0,0006)	-0,0004 (-0,0003)	0,0011 (0,0010)	0,0009 (0,0009)
Biais absolu de $\hat{p}_{im+}$	0,0076 (0,0078)	0,0051 (0,0055)	0,0089 (0,0085)	0,0048 (0,0046)	0,0108 (0,0106)	0,0074 (0,0073)
REQM empirique	0,0479 (0,0483)	0,0467 (0,0469)	0,0417 (0,0414)	0,0401 (0,0402)	0,0236 (0,0233)	0,0231 (0,0229)
Biais relatif de $\sqrt{\widehat{\mathrm{Var}}(\hat{p}_{im+})}$	-0,1192 (-0,1197)	-0,1125 (-0,1128)	-0,1273 (-0,1276)	-0,1180 (-0,1186)	-0,1524 (-0,1521)	-0,1376 (-0,1372)
Biais relatif absolu de $\sqrt{\widehat{\mathrm{Var}}(\hat{p}_{im+})}$	0,1192 (0,1197)	0,1125 (0,1128)	0,1273 (0,1276)	0,1180 (0,1186)	0,1524 (0,1521)	0,1376 (0,1372)
Biais relatif de $\sqrt{\widehat{\mathrm{Var}}^{(B)}(\hat{p}_{im+})}$	-0,0275 (-0,0272)	-0,0173 (-0,0175)	-0,0309 (-0,0314)	-0,0204 (-0,0207)	-0,0391 (-0,0393)	-0,0273 (-0,0269)
Biais relatif absolu de $\sqrt{\widehat{\mathrm{Var}}^{(B)}(\hat{p}_{im+})}$	0,0294 (0,0290)	0,0227 (0,0228)	0,0349 (0,0343)	0,0263 (0,0265)	0,0450 (0,0446)	0,0353 (0,0347)
Taux de couverture naïf	91,35 (91,325)	91,91 (91,875)	91,19 (91,225)	91,78 (91,750)	90,67 (90,650)	91,26 (91,300)
Écart absolu de la couverture naïf par rapport au taux nominal de 95 %	3,65 (3,675)	3,09 (3,125)	3,81 (3,775)	3,22 (3,250)	4,33 (4,350)	3,74 (3,700)
Taux de couverture ajusté	94,44 (94,400)	94,75 (94,775)	94,37 (94,350)	94,68 (94,650)	93,91 (93,925)	94,40 (94,375)
Écart absolu de la couverture ajusté par rapport au taux nominal de 95 %	1,58 (1,600)	1,43 (1,425)	1,71 (1,725)	1,50 (1,525)	1,91 (1,900)	1,62 (1,650)

ordinal de la catégorie de revenu conditionnel à un revenu différent de zéro.

En pratique, les données historiques sont souvent disponibles aux fins de la planification des enquêtes. Par exemple, la sélection des variables aux fins des prédictions du modèle pourrait être fondée sur les données des recensements antérieurs. Pour simuler cette situation, un échantillon aléatoire de 2 000 sujets a été tiré de l'échantillon de 1 %. Les variables pour la prédiction du modèle ont été déterminées en appliquant une méthode de régression logistique par degrés; il s'agissait de l'âge, du sexe et de la race (blancs, noirs ou autres).

Ainsi, les modèles multinomial et ordinal utilisés dans la présente étude incluaient quatre variables individuelles explicatives pour l'âge, le sexe et la race (deux variables indicatrices étaient requises pour coder les diverses races). Toutefois, ils comprenaient également quatre variables de la région locale représentant l'âge moyen, la proportion d'hommes, la proportion de blancs et la proportion de noirs. Peu importe le modèle retenu, ces variables au niveau de la région locale sont nécessaires. En effet, lorsqu'elles sont exclues, on observe que plus la valeur attendue de $\hat{p}_{im+}$ augmente, plus le biais augmente également, passant d'une grande valeur négative à une grande valeur positive. L'inclusion de covariables appartenant au niveau du domaine élimine cette corrélation. En conséquence, puisque les variables de la région locale sont également incluses dans les modèles, le modèle multinomial contient 18 paramètres des effets fixes (deux pour chacun des niveaux individuels et des variables explicatives de la région locale, et deux termes constants) et 40 effets aléatoires (deux pour chacune des 20 régions locales échantillonnées), tandis que le modèle ordinal contient dix paramètres des effets fixes (un pour chacune des variables explicative des niveaux individuel et régional et deux termes constants) et 40 effets aléatoires (deux pour chacune des 20 régions locales échantillonnées). Pour une étude détaillée comparant les modèles de régression logistique pour l'estimation des proportions pour petites régions avec ou sans covariables du domaine et qui utilisent des données binômiales, voir Farrell et coll. (1997a).

Les données servant à l'estimation des proportions de sujets de chacune des régions locales appartenant aux diverses catégories de revenu ont été obtenues à partir de l'échantillon de 1 % à l'aide d'un plan d'échantillonnage autopondéré à deux degrés. Au premier degré, 20 des 42 régions locales ont été choisies sans remise, avec probabilité proportionnelle à la taille (PPT). La méthode utilisée pour choisir ces régions locales était en fait la méthode d'échantillonnage systématique aléatoire avec PPT (voir Kish 1965, p. 230). À la seconde étape, 50 sujets ont été choisis au hasard dans chacune des régions locales. Cinq cent échantillons ont ainsi été tirés à l'aide de ce plan à deux degrés. Toutefois, on n'a pas procédé à l'échantillonnage répété au stade de la sélection des régions locales. Ainsi, les 500 échantillons ont été tirés des 20 mêmes régions locales. Pour ces 20 régions, les proportions moyennes par région locale pour les catégories 1, 2 et 3 des niveaux de revenu sont 0,7142, 0,2260 et 0,0598.

Il convient de noter que pour le modèle ordinal, la contrainte $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ doit être respectée pour que $\pi_{ij2} \geq 0$. Une vérification de cette contrainte pour chacun des 500 échantillons à l'aide des estimations des termes constants et des effets aléatoires a permis de vérifier qu'elle était respectée dans tous les cas. En fait, on a découvert que pour chacun des 500 échantillons tirés, la différence observée dans les estimations des termes constants était toujours positive, supérieure d'au moins deux ordres de grandeur à la majorité des différences absolues des estimations des effets aléatoires, et supérieure d'un ordre de grandeur à la totalité d'entre elles. Ainsi, les termes constants dans le modèle dominent les effets aléatoires.

Pour comparer les propriétés des estimateurs des proportions dans les petites régions sur une série de répétitions du plan de sondage, pour chacun des 500 échantillons choisis, les valeurs $\hat{p}_{im+}$, $\widehat{\mathrm{Var}}(\hat{p}_{im+})$, et $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ associées à chaque niveau de revenu ont été obtenues pour chaque région locale, échantillonnée ou non, à l'aide des modèles multinomial et ordinal. Pour chaque modèle, les estimations de $\widehat{\mathrm{Var}}(\hat{p}_{im+})$ et $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ ont servi respectivement à établir des intervalles de confiance symétriques (95 %), naïfs ou ajustés par la méthode bootstrap, de la distribution empirique de Bayes. Les estimations de $\widehat{\mathrm{Var}}^{\{B\}}(\hat{p}_{im+})$ ont été obtenues à l'aide de la méthode bootstrap afin de générer 100 échantillons bootstrap pour chacun des 500 échantillons de simulation.

Il convient de noter que pour le modèle ordinal, la contrainte $\beta_{02} - \beta_{01} \geq \delta_{i1} - \delta_{i2}$ doit également être respectée dans la méthode bootstrap pour les effets aléatoires générés à partir d'une distribution estimée; la création des échantillons bootstrap risquerait autrement de donner des estimations négatives pour certaines des probabilités π_{ijm+}. Tout au long de la simulation pour l'application examinée ici, aucune probabilité négative n'a été relevée lors de l'utilisation de la méthode bootstrap. Une des méthodes envisageables pour l'évaluation de la vraisemblance des probabilités négatives pendant l'application de la méthode bootstrap consiste à considérer le rapport de la différence $\hat{\beta}_{02} - \hat{\beta}_{01}$ sur l'écart-type antérieur estimé de la différence $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. Ce rapport a été déterminé pour chaque région locale échantillonnée dans chacun des 500 échantillons de simulation tirés. La moyenne de ce groupe entier de rapports était 6,8, et aucun n'était inférieur à 5,8. Ainsi, on a déterminé que la différence $\hat{\beta}_{02} - \hat{\beta}_{01}$ était toujours au moins 5,8 fois plus grande que l'écart-type estimé de la différence $\hat{\delta}_{i1} - \hat{\delta}_{i2}$. On pourrait ainsi empiriquement conclure que lorsque le ratio décrit ci-dessus est d'au moins 3, il est hautement improbable que la méthode bootstrap conduise à des probabilités négatives.

Nous présentons au tableau 1 les statistiques sommaires moyennes des 500 échantillons de simulations obtenus pour les modèles multinomial et ordinal sur l'ensemble des régions locales échantillonnées pour chacune des trois catégories de revenu. Une étude de la stabilité de ces statistiques a été réalisée en examinant comment elle changeait sous l'effet de la prise d'échantillons supplémentaires. Seuls des changements minimes ont été observés

La méthode décrite dans la présente section peut également servir à élaborer des estimations ponctuelles et des estimations d'intervalles pour les proportions de petites régions fondées sur $\hat{p}_{im+}$ et $\tilde{p}_{im+}$, lorsqu'on utilise un modèle ordinal. Dans la présente étude, nous proposons un modèle à effets fixes et aléatoires pour la valeur de π_{ijm+} fondée sur le modèle ordinal proposé par McCullagh (1980):

$$\log\left(\frac{\pi_{ij1} + \ldots + \pi_{ijm}}{\pi_{ij(m+1)} + \ldots + \pi_{ijm}}\right) = \beta_{0m} - \underline{X}_{ij}^T \underline{\beta} + \delta_{im}, \tag{2.5}$$

$$\underline{\delta}_i \sim \text{ i.i.d. Normale } (\underline{0}, \boldsymbol{D}).$$

Le vecteur $\underline{X}_{ij}$ contient les valeurs des variables explicatives des effets fixes pour le ij-ième sujet, tandis que $\underline{\beta}$ représente un vecteur des paramètres des effets fixes. Il existe un terme constant, β_{0m}, qui est associé à la m-ième catégorie de variables de réponses. On présume ici encore que les effets aléatoires ont une distribution normale. Il convient de noter que le modèle (2.5) exige en particulier que la restriction $\beta_{0(m+1)} - \beta_{0m} \geq \delta_{im} - \delta_{i(m+1)}$ se réalise pour que $\pi_{ij(m+1)} \geq 0$. Nous revenons en détails sur cette contrainte à la section 3.

La démarche choisie pour donner l'approximation de l'incertitude en $\hat{p}_{im+}$ et $\tilde{p}_{im+}$ lorsque π_{ijm+} est fondé sur la formule (2.3) ou (2.5) peut être qualifiée de naïve, puisque $\widehat{\text{Var}}(\hat{p}_{im+})$ et $\widehat{\text{Var}}(\tilde{p}_{im+})$ ne tiennent pas compte de l'incertitude qui découle de l'estimation des paramètres de la distribution des effets aléatoires. Ainsi, les estimations d'intervalles pour p_{im+} qui sont fondées sur $\widehat{\text{Var}}(\hat{p}_{im+})$ et $\widehat{\text{Var}}(\tilde{p}_{im+})$ sont typiquement trop courtes. On a proposé de nombreuses méthodes pour corriger ce problème (voir Carlin et Gelfand 1990; et Laird et Louis 1987). Dans la présente étude, la méthode bootstrap de type III proposée par Laird et Louis (1987) sert à ajuster les mesures d'incertitude obtenues par l'estimation naïve. Cette méthode est décrite par Farrell et coll. (1997a), pour une variable de résultats binômiale. Elle peut être adaptée à (2.3) ou à (2.5), et s'applique peu importe que l'estimation soit fondée sur $\hat{p}_{im+}$ ou sur $\tilde{p}_{im+}$.

La méthode exige qu'un certain nombre d'échantillons bootstrap, N_B, soient générés pour un ensemble particulier de données. Supposons que l'estimation de la proportion pour la petite région doive être fondée sur $\hat{p}_{im+}$. Pour le b-ième échantillon bootstrap, on obtient une estimation $\hat{p}_{bim+}$ pour p_{im+} fondée sur (2.3) ou (2.5) en même temps qu'une estimation naïve de la variabilité de $\hat{p}_{bim+}$, $\widehat{\text{Var}}(\hat{p}_{bim+})$. Les valeurs $\hat{p}_{bim+}$ et $\widehat{\text{Var}}(\hat{p}_{bim+})$ sont déterminées pour chacun des N_B échantillons bootstrap, et servent à calculer une estimation de la variabilité associée à $\hat{p}_{im+}$ et ajustée selon la méthode bootstrap:

$$\widehat{\text{Var}}^{\{B\}}(\hat{p}_{im+}) = \frac{\sum_b \widehat{\text{Var}}(\hat{p}_{bim+})}{N_B} + \frac{\sum_b (\hat{p}_{bim+} - \hat{p}_{im+}^{\{B\}})^2}{N_B - 1},$$

où

$$\hat{p}_{im+}^{\{B\}} = \frac{\sum_b \hat{p}_{bim+}}{N_B}.$$

Il convient de noter que même si les sujets ne sont pas choisis par échantillonnage aléatoire simple sans remise, dans la présente étude, les données de sondage n'ont pas été pondérées. Toutefois, en pratique, les poids attachés à un enregistrement varieront en fonction de caractéristiques du plan de sondage telles que la non-réponse différentielle et la répartition en grappes. Dans la présente étude, les modèles tiennent compte des effets de ces caractéristiques. De plus amples recherches seront nécessaires pour déterminer qu'elles sont les incidences sur la méthode bootstrap de l'incorporation dans les modèles de poids liés aux sondages.

3. EXEMPLE PRATIQUE

On a procédé à une comparaison des estimations de proportions pour petites régions fondée sur des modèles logistiques multinomial ou ordinal en utilisant une étude de simulation où la variable de réponses était ordinale. L'ensemble de données est fondé sur un échantillon de 1 % prélevé à même le recensement américain de 1950 (United States Bureau of the Census 1984). On utilise les données fondées sur le recensement de 1950 puisqu'il s'agit d'un échantillon de micro-données accessible au public et que aucun des recensements plus récents n'est disponible sous cette forme. Ainsi, les résultats examinés ci-après pour les modèles multinomial ou ordinal sont obtenus en utilisant des variables explicatives pour chaque sujet à l'intérieur d'une région locale. Pour un examen détaillé des difficultés rencontrées dans la recherche des micro-données, voir Bethlehem, Keller et Pannekoek (1990).

L'application envisagée est l'estimation de la proportion des personnes vivant dans une région locale donnée correspondant à chacune des trois catégories de variables de résultats ordinales représentant le revenu personnel total, où la région locale correspond typiquement à un État. Cette variable englobe toutes les sources de revenu, y compris les salaires, les revenus d'affaires et les revenus nets provenant d'autres sources. Les catégories utilisées sont celles des personnes à faible revenu (moins de 2 500$), à revenu moyen (2 500$ à moins de 10 000$) ou à revenu élevé (10 000$ et plus) en 1949. Ainsi, $m = 1$ pour les personnes à faible revenu (catégorie 1), $m = 2$ pour les personnes à revenu moyen (catégorie 2) et $m = 3$ pour les personnes à revenu élevé (catégorie 3). Les modèles multinomial et ordinal ont chacun été utilisés pour obtenir des estimations ponctuelles et des estimations d'intervalles dans 42 régions locales. Vingt de ces régions ont été échantillonnées. Il convient de noter que les personnes sans revenu ont été incluses dans la catégorie 1. On aurait pu, en guise de solution de rechange, procéder en deux étapes: d'abord avec un modèle logistique de la probabilité d'un revenu différent de zéro, et ensuite avec un modèle multinomial ou

Pour obtenir les estimations de Bayes des paramètres du modèle, on attribue des valeurs quelconques aux paramètres inconnus de la distribution des effets aléatoires. Désignons par $\underline{y}_{ij}^T = (y_{ij1}, ..., y_{ijM})$ un vecteur du ij-ième sujet échantillonné où la composante associée à la catégorie de variable de résultats à laquelle cette personne appartient a une valeur de un. Les entrées qui restent sont égales à zéro. Si $\boldsymbol{Y}$ est une matrice dont les rangs sont désignés par $\underline{y}_{ij}^T$, les données seront alors distribuées comme suit:

$$f(\boldsymbol{Y} \mid \underline{\beta}, \underline{\delta}_c) \,\alpha \prod_{ij} \pi_{ij1}^{y_{ij1}} \pi_{ij2}^{y_{ij2}} \ldots \pi_{ijM}^{y_{ijM}},$$

où $\underline{\beta}^T = (\underline{\beta}_1^T, ..., \underline{\beta}_{M-1}^T)$, et $\underline{\delta}_c^T = (\underline{\delta}_1^T, ..., \underline{\delta}_I^T)$. Si une distribution uniforme est précisée pour les effets fixes, la distribution des paramètres devient $f(\underline{\beta}, \underline{\delta}_c | \boldsymbol{D}_c) \propto \exp(-½\underline{\delta}_c^T \boldsymbol{D}_c \underline{\delta}_c)$, où $\boldsymbol{D}_c = \operatorname{diag}(\boldsymbol{D}, \boldsymbol{D}, ..., \boldsymbol{D})$. La distribution combinée des données et des paramètres est déterminée en utilisant $f(\boldsymbol{Y} | \underline{\beta}, \underline{\delta}_c)$ et $f(\underline{\beta}, \underline{\delta}_c | \boldsymbol{D}_c)$, et utilisée par la suite pour obtenir la distribution postérieure des paramètres. Malheureusement, il est impossible de dériver une forme fermée de cette distribution postérieure à cause du caractère insoluble de l'intégration requise pour obtenir la distribution marginale de $\boldsymbol{Y}$. Une méthode d'intégration stochastique comme celle de l'échantillonnage de Gibbs (voir Zeger et Karim 1991) représenterait une solution possible. Ripley et Kirkland (1990) indiquent qu'une telle démarche présenterait notamment l'inconvénient de nécessiter des calculs intensifs et de laisser planer des incertitudes quant au moment où le processus d'échantillonnage parvient à l'équilibre. Comme le temps de calcul est une préoccupation particulière de la simulation examinée à la section 3, nous ne nous y attarderons pas plus avant ici. Par ailleurs, Breslow et Clayton (1993) mentionnent qu'on peut toujours envisager des méthodes simples et approximatives. Beaucoup de chercheurs ont démontré qu'une approximation normale multivariée de la distribution postérieure donne d'excellents résultats en pratique (voir Farrell et coll. 1997a; Laird 1978; Tomberlin 1988 et Wong et Mason 1985). Breslow et Lin (1995) rappellent toutefois qu'une telle méthode pourrait donner des estimations incohérentes pour les paramètres à effets fixes. Ainsi, si $\hat{p}_{im+}$ doit être fondé sur des estimations des effets fixes obtenues de cette façon, la même mise en garde risquera de s'appliquer en ce qui à trait à la cohérence de $\hat{p}_{im+}$ pour l'estimation de p_{im+}.

Selon Farrell et coll. (1997a), une distribution normale multivariée dont la moyenne correspond au mode et dont la matrice de covariance est égale à l'inverse de la matrice d'information évaluée au mode représente une approximation de la distribution postérieure des paramètres. La matrice d'information dont il est question ici est simplement la deuxième dérivée de la distribution postérieure calculée par rapport à $\underline{\beta}$ et à $\underline{\delta}_c$. Lorsque des valeurs sont précisées pour les paramètres inconnus de la distribution des effets aléatoires, le mode et la matrice de covariance qui en découlent constituent un ensemble initial d'estimations des paramètres du modèle. Les estimations empiriques de Bayes sont alors obtenues en utilisant l'algorithme EM décrit par Dempster, Laird et Rubin (1977) afin de déterminer les estimations des paramètres de la distribution des effets aléatoires. L'algorithme converge rapidement, en quelques minutes seulement en temps réel. Pour en savoir plus sur la façon d'obtenir les estimations empiriques de Bayes pour un modèle fondé sur un plan d'échantillonnage à deux degrés et une variable de réponses binômiale, voir MacGibbon et Tomberlin (1989).

Les estimations empiriques de Bayes des paramètres du modèle sont utilisés en (2.2) pour déterminer $\hat{p}_{im+}$. Pour élaborer une expression correspondant à l'incertitude de $\hat{p}_{im+}$, on présume que la valeur N_i est connue. Comme la démarche utilisée est fondée sur un modèle et qu'elle est prédictive par nature, l'incertitude entourant $\hat{p}_{im+}$ découle uniquement du terme $\sum \hat{\pi}_{ijm+}$; le terme $\sum y_{ijm+}$ a une variance de zéro. Ainsi, l'erreur quadratique moyenne de $\hat{p}_{im+}$ en tant que prédicteur de p_{im+} peut être estimée comme suit:

$$\widehat{\text{REQM}}(\hat{p}_{im^+}) = \widehat{\text{Var}}\left(\frac{\sum_{j \in S'} \hat{\pi}_{ijm^+}}{N_i}\right) + \frac{\sum_{j \in S'} \hat{\pi}_{ijm^+}(1 - \hat{\pi}_{ijm^+})}{N_i^2}. \qquad (2.4)$$

Pour les régions locales échantillonnées où n_i est plus grand que zéro, le premier terme de (2.4) est de l'ordre de $1/n_i$, tandis que le second est de l'ordre de $1/N_i$. Dans cette étude, l'approximation de l'erreur quadratique moyenne de $\hat{p}_{im+}$ est fondée sur le premier terme uniquement, lequel donne une approximation utile à condition que N_i soit grand comparativement à n_i. Pour les régions locales non échantillonnées, le premier terme de (2.4) est de l'ordre de 1; il domine donc toujours le second terme.

Pour estimer l'incertitude de $\hat{p}_{im+}$, qui est exprimée sous forme de fonction non linéaire des estimateurs des effets fixes et aléatoires, l'expression de $\hat{p}_{im+}$ est linéarisée par développement en une série de Taylor multivariée de premier ordre autour des valeurs réalisées des effets fixes et aléatoires. La variance de l'expression qui en découle, désignée par $\widehat{\text{Var}}(\hat{p}_{im+})$, est assimilée à une estimation de l'incertitude de $\hat{p}_{im+}$. Farrell et coll. (1997a), fournissent des informations détaillées sur le développement des séries de Taylor pour une variable de résultats binômiale.

Lorsque les micro-données de population pour les variables auxiliaires ne sont pas disponibles, il est impossible de déterminer $\hat{p}_{im+}$ avec (2.2). Pour les modèles non linéaires comme (2.3), la prédiction n'est pas directe dans une telle situation. Toutefois, un estimateur de rechange de $\hat{p}_{im+}$, $\tilde{p}_{im+}$ par exemple, qui nécessite uniquement des statistiques sommaires de la région locale (un vecteur de la moyenne et une matrice de covariances de la population finie) tant pour les variables continues que pour les variables nominales peut être obtenu en adaptant la démarche proposée par Farrell, MacGibbon et Tomberlin (1997b) aux fins de la réalisation de cet objectif lorsqu'on cherche à estimer les paramètres binômiaux des petites régions. Ce même développement des séries de Taylor qu'on a utilisé pour estimer l'exactitude de $\hat{p}_{im+}$ peut être employée pour obtenir une mesure de l'incertitude dans le cas de $\tilde{p}_{im+}$, $\widehat{\text{Var}}(\tilde{p}_{im+})$.

estimations des proportions pour petites régions fondées sur une variable ordinale en utilisant un modèle multinomial ou ordinal, nous appliquons les méthodes empiriques proposées par Bayes à des données issues du recensement américain de 1950 afin de prédire, pour une petite région donnée, la proportion des personnes appartenant aux diverses catégories d'une variable de réponses ordinale représentant le niveau de revenu.

Ce genre d'estimation pose de nombreux problèmes sur lesquels il convient de se pencher. On peut mentionner en particulier la sélection des variables explicatives pour le modèle, les diagnostics du modèle, le plan de sondage et les propriétés des estimateurs utilisées. Par exemple, parmi les diagnostics pour les modèles multinomial et ordinal figurait une évaluation de l'ajustement du modèle fondée sur les valeurs. Farrell (1991) a proposé une description de ce diagnostic et d'autres diagnostics. Les résultats ne semblaient pas indiquer une absence d'ajustement pour l'un ou l'autre des modèles. Dans la présente étude, nous cherchons surtout à déterminer les propriétés des estimateurs empiriques de Bayes pendant l'utilisation répétée du plan de sondage à l'aide d'une simulation. Pour de nombreux spécialistes d'enquêtes, de telles propriétés revêtent une importance primordiale.

On reproche notamment aux méthodes empiriques de Bayes d'utiliser des estimations d'intervalles qui ne donnent pas le niveau souhaité de couverture puisque l'incertitude qui découle de l'obligation d'estimer les paramètres de la distribution antérieure n'est pas prise en compte. Dans la présente étude, nous avons recours comme le suggèrent Laird et Louis (1987) aux méthodes bootstrap pour l'ajustement d'estimations naïves de l'exactitude. Par ailleurs, Prasad et Rao (1990) ont mis au point une méthode qui tente de «capturer» l'incertitude qui n'est pas prise en compte par les estimations naïves. Cette méthode a été conçue pour trois modèles linéaires spécifiques contenant des effets aléatoires, mais Cressie (1992) a déterminé certaines situations où elle pourrait être approprié. Il importe en particulier de souligner que les résultats obtenus doivent obéir à une distribution normale.

L'estimation par des méthodes empiriques de Bayes fondée sur un modèle logistique multinomial ou ordinal est décrite à la section 2. L'étude de simulation visant à comparer les modèles logistiques multinomial et ordinal pour les réponses ordinales est décrite à la section 3. Nos observations et nos conclusions sont présentées à la section 4.

2. MÉTHODES D'ESTIMATION

Imaginons une caractéristique d'intérêt pour une petite région discrète comportant M résultats possibles. L'indice m permet d'identifier les catégories, où $m = 1, \ldots, M-1$ et $m^+ = 1, \ldots, M$. En outre, les lettres minuscules et majuscules soulignées désignent des vecteurs tandis que les lettres majuscules en caractères gras représentent des matrices.

Les méthodes d'estimation sont illustrées dans un plan d'échantillonnage à deux degrés où les sujets sont choisis à partir de régions locales présélectionnées. Ainsi, les régions locales constituent les primaires unités d'échantillonnage. Désignons par p_{im^+} la proportion des personnes vivant dans la i-ième région locale qui appartiennent à la catégorie m^+ de la variable de réponses. On obtient alors

$$p_{im^+} = \sum_j y_{ijm^+}/N_i, \tag{2.1}$$

où y_{ijm^+} est égal à 0 ou à 1, selon que la j-ième personne de la région locale i appartient à la catégorie m^+ de la caractéristique d'intérêt et N_i désigne la taille de la population de la i-ième région locale.

La méthode utilisée par Farrell et coll. (1997a), pour estimer les proportions pour petites régions en se fondant sur les variables de résultat binômiales est adaptée ici pour permettre l'estimation de p_{im^+}. Cette méthode s'inspire de la démarche explicitement fondée sur la modélisation proposée par Dempster et Tomberlin (1980). Désignons par π_{ijm^+} la probabilité que la j-ième personne appartenant à la i-ième région locale appartienne à la catégorie m^+ de la variable de réponses. Dans ce cas, selon Royall (1970), la valeur p_{im^+} de l'équation (2.1) est estimée par

$$\hat{p}_{im^+} = \left(\sum_{j \in S} y_{ijm^+} + \sum_{j \in S'} \hat{\pi}_{ijm^+} \right) \Big/ N_i, \tag{2.2}$$

où S représente l'ensemble des n_i personnes échantillonnées dans la région locale i, et S' désigne l'ensemble des personnes appartenant à la région locale i non incluses dans l'échantillon. Il nous reste maintenant à déterminer les valeurs de $\hat{\pi}_{ijm^+}$. Pour obtenir ces estimations, on utilise des modèles de régression logistique afin de décrire les probabilités associées aux membres de la population.

Dans un modèle logistique multinomial, les valeurs π_{ijm^+} sont décrites comme suit:

$$\begin{aligned} \log(\pi_{ijm} / \pi_{ijM}) &= \underline{X}_{ij}^T \underline{\beta}_m + \delta_{im}, \\ \underline{\delta}_i &\sim \text{i.i.d. Normal}\,(\underline{0}, \boldsymbol{D}), \end{aligned} \tag{2.3}$$

où $\underline{\delta}_i^T = (\delta_{i1}, \ldots, \delta_{i(M-1)})$, $i = 1, \ldots, I$, et $\boldsymbol{D}$ désigne une matrice de covariance inconnue. Dans ce modèle, $\underline{X}_{ij}$ est un vecteur des variables explicatives à effets fixes, le vecteur $\underline{\beta}_m$ contient les paramètres à effets fixes associés à la m-ième catégorie de la variable d'intérêt et δ_{im} désigne un effet aléatoire à distribution normale associé à la m-ième catégorie de la caractéristique d'intérêt dans la i-ième région locale. Le vecteur $\underline{X}_{ij}$ peut inclure des covariables tant au niveau individuel qu'au niveau agrégé. Pour les plans de sondage comportant plus de deux étapes, un modèle analogue contiendrait les effets aléatoires pour les unités d'échantillonnage à chaque stade, à l'exclusion du stade final.

À noter que le modèle indiqué en (2.3), contrairement à un modèle semblable proposé par Malec et coll. (1993), ne contient pas de termes d'interaction entre les effets de la région locale et les variables explicatives à effets fixes. Toutefois, les termes permettant de tenir compte d'une telle interaction seraient inclus s'ils étaient jugés nécessaires.

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, p. 127-134
Statistique Canada

Estimation de proportions pour petites régions par des méthodes empiriques de Bayes, à partir de variables ordinales

PATRICK J. FARRELL[1]

RÉSUMÉ

La modélisation des réponses ordinales a déjà fait l'objet de beaucoup de recherches. Selon certains auteurs, lorsque la variable de réponses est ordinale, la prise en compte de cette caractéristique dans le modèle à estimer devrait accroître la performance de ce modèle. Dans des conditions ordinales, Campbell et Donner (1989) ont comparé le taux asymptotique d'erreurs de classification du modèle logistique multinomial à celui du modèle logistique ordinal d'Anderson (1984). Ils ont démontré que ce dernier était assorti d'un taux asymptotique d'erreurs prévisible inférieur à celui du modèle logistique multinomial. Dans le présent article, nous cherchons à comparer la performance d'un modèle logistique ordinal et d'un modèle multinomial pour les réponses ordinales. Toutefois, au lieu de concentrer notre attention sur l'efficacité de classification, nous nous attachons à estimer les proportions pour les petites régions. En utilisant un modèle logistique multinomial et un modèle ordinal, nous cherchons plus particulièrement à adapter l'estimation de proportions pour petites régions à partir de données binômiales par des méthodes empiriques de Bayes, tel que le suggèrent Farrell, MacGibbon et Tomberlin (1997a), aux variables qui appartiennent à plus de deux catégories de résultats. Les propriétés des estimateurs fondés sur ces deux modèles sont comparées au moyen d'une simulation au cours de laquelle les méthodes empiriques de Bayes proposées sont appliquées à des données issues du recensement américain de 1950, afin de chercher à prévoir, pour des petites régions, les proportions des personnes appartenant aux diverses catégories d'une variable de réponses ordinale représentant le niveau de revenu.

MOTS CLÉS: Méthode bootstrap; plan d'enquête complexe; régression logistique; modèles d'effets aléatoires; statistiques sommaires sur les petites régions; séries de Taylor.

1. INTRODUCTION

La modélisation des réponses ordinales a fait l'objet de beaucoup de recherches (voir Albert et Chib 1993; Anderson 1984; Crouchley 1995 et McCullagh 1980). Selon certains auteurs, lorsque la variable de réponses est ordinale, la prise en compte de cette caractéristique dans le modèle à estimer devrait améliorer la performance de ce modèle. Dans des conditions ordinales, Campbell et Donner (1989) ont comparé théoriquement le taux asymptotique d'erreurs de classification du modèle logistique multinomial à celui du modèle logistique ordinal d'Anderson (1984), démontrant que le modèle ordinal présentait un taux asymptotique d'erreurs prévisible plus bas. Toutefois, dans une simulation subséquente, Campbell, Donner et Webster (1991) ont démontré que les modèles ordinaux donnent une classification moins exacte que les modèles multinomiaux dans toutes sortes de circonstances; ils en ont conclu que ces modèles ne présentent aucun avantage lorsque la classification constitue le principal objectif de l'analyse.

Nous cherchons également, dans le présent article, à comparer la performance d'un modèle logistique ordinal et d'un modèle multinomial pour les réponses ordinales. Toutefois, au lieu de concentrer notre attention sur l'efficacité de la classification, nous nous attachons à estimer les proportions pour les petites régions.

L'estimation des paramètres d'une petite région est un problème d'échantillonnage d'une population finie qui a déjà fait l'objet d'énormément d'attention. Ghosh et Rao (1994) proposent un excellent tour d'horizon de ces recherches. Ils démontrent que lorsqu'on les utilise en guise de solution de compromis entre l'estimateur synthétique et l'estimateur direct, les estimateurs fondés sur les méthodes empiriques ou hiérarchiques de Bayes ne sont pas exposés aux biais importants parfois associés à l'estimateur synthétique (voir Gonzales 1973); ils ne sont pas non plus aussi variables qu'un estimateur direct. Farrell, MacGibbon et Tomberlin (1997a) arrivent à une conclusion semblable à la suite d'une étude des méthodes empiriques de Bayes pour l'estimation de proportions pour une petite région à partir d'une variable de résultats binômiale.

Malgré les nombreux travaux qui ont cherché à prévoir les proportions pour petites régions à partir de variables de réponses binômiales (voir Dempster et Tomberlin 1980; MacGibbon et Tomberlin 1989; Farrell 1991; Farrell et coll. 1997a; Malec, Sedransk et Tompkins 1993; Stroud 1991 et Wong et Mason 1985), on s'est très peu intéressé à l'estimation des proportions fondées sur les variables de réponses appartenant à plus de deux catégories de résultats. Dans le présent article, nous adaptons la démarche empirique de Bayes utilisée par Farrell et coll. (1997a), à de telles variables en fondant nos estimations sur des modèles logistiques multinomial ou ordinal. Pour comparer les

[1] Patrick J. Farrell, professeur adjoint, Department of Mathematics and Statistics, Acadia University, Wolfville, (Nouvelle-Écosse), B0P 1X0.

BIBLIOGRAPHIE

BREWER, K.R.W., EARLY, L.J., et HANIF M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.

COTTON, F., et HESSE C. (1992). Tirages coordonnés d'échantillons. Document de travail E9206 de l'INSEE.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

HIDIROGLOU, M.A., CHOUDHRY, G.H., et LAVALLÉE, P. (1991). Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises. *Techniques d'enquête*, 17, 221-227.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

et c'est même plus simple. Le plus délicat est le retirage après la phase intermédiaire de rotation. Non seulement il s'agit d'obtenir un TASST mais aussi d'avoir, si possible, le même recouvrement que pour la méthode 1 de la section 5.

Posons $\alpha_{\mathbf{h}_1}(j)$ le numéro ω de l'unité de rang j dans une ancienne strate $\mathbf{h}_1$.

Supposons d'abord que, dans une ancienne strate, toutes les unités soient telles que $f_{h_2} \geq n'_{\mathbf{h}_1}/N_{\mathbf{h}_1}$. En particulier cela se produit dans toutes les strates pour un sondage avec un seul taux dans la partie sondée, si on ne baisse pas ce taux. On cherche alors une transformation telle que les numéros des unités de l'échantillon se retrouvent au début de $[0,1)$. La plus simple est la permutation:

$$\begin{cases} \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j + N_{\mathbf{h}_1} - R_{\mathbf{h}_1,d}), & j \leq R_{\mathbf{h}_1,d}, \\ \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j - R_{\mathbf{h}_1,d}), & j > R_{\mathbf{h}_1,d}. \end{cases}$$

Cependant une transformation moins coûteuse est:

$$\begin{cases} \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j) + \alpha_{\mathbf{h}_1}(N_{\mathbf{h}_1}) - \alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d}), & j \leq R_{\mathbf{h}_1,d}, \\ \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j) - \alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d}), & j > R_{\mathbf{h}_1,d}. \end{cases}$$

Il suffit d'aller rechercher $\alpha_{\mathbf{h}_1}(R_{\mathbf{h}_1,d})$ et $\alpha_{\mathbf{h}_1}(N_{\mathbf{h}_1})$, après quoi un simple calcul séquentiel permet de déduire β de α.

Le Jacobien de la transformation est égal à 1 et par conséquent les numéros conservent leur distribution uniforme. Par ailleurs la loi conjointe $p(s_1, s_2)$ est la même que s'il n'y avait pas eu rotation. La démonstration figure dans Cotton et Hesse (1992, page 55). On a donc le recouvrement maximum de TASST.

Si dans la strate on a des unités avec $f_{h_2} < n'_{\mathbf{h}_1}/N_{\mathbf{h}_1}$ et qu'on applique la transformation, les unités dont le rang est, en gros, compris entre $N_{\mathbf{h}_1} f_{h_2}$ et $n'_{\mathbf{h}_1}$ ne sont pas reprises lors du retirage mais vont être réintroduites à l'occasion d'une prochaine rotation. Il est donc préférable d'utiliser pour ces unités une transformation qui situe juste avant f_{h_2} les nouveaux numéros. On doit procéder par sous-ensembles selon la valeur de f_{h_2}. Mais cela tend à diminuer le recouvrement.

REMERCIEMENTS

Le point de départ de nos réflexions est un document interne de la Division des méthodes d'enquêtes-entreprises à Statistique Canada: Hidiroglou M.A., Srinath K.P. (1990), Methods of integrated sampling for sub-annual business surveys.

Nous remercions un rédacteur associé et un arbitre anonymes pour leur aide apportée à la rédaction de cet article.

Certaines des méthodes proposées ont été appliquées à l'INSEE, mais les opinions exprimées n'engagent que les auteurs.

ANNEXE

Les probabilités d'inclusion du premier ordre dans la méthode de Kish et Scott (1971)

Donnons un exemple où la probabilité d'inclusion du premier ordre n'est pas strictement contrôlée.

La population est divisée en trois parties A, B et C d'égale taille N. Le premier tirage est un TAS de $2a$ unités dans $A + B$ et un TAS de a unités dans C. Au deuxième tirage, on veut tirer a unités dans A et $2a$ unités dans $B + C$, en retenant le maximum d'unités du premier échantillon et avec la probabilité d'inclusion uniforme a/N. La méthode de Kish et Scott (1971) consiste à rajouter ou retrancher par TAS le nombre convenable d'unités séparément dans A et dans $B + C$. Dans A, le second tirage marginal est un TAS et la probabilité d'inclusion est bien uniforme. Montrons qu'il n'en est pas de même dans $B + C$. Soient n_1 et n_2 les tailles des deux échantillons successifs dans B. Par symétrie, la probabilité d'inclusion au second tirage est uniforme dans B. Elle y vaut:

$$E(n_2)/N = [E(n_1) + E(n_2 - n_1)]/N$$
$$= a/N + E(n_2 - n_1)/N.$$

Si $n_1 = a, n_2 - n_1 = 0$; sinon l'espérance de $n_2 - n_1$ conditionnelle à n_1 diffère selon le signe de $a - n_1$:

Si $a - n_1 > 0, E[(n_2 - n_1) \mid n_1] = (a - n_1)(N - n_1)/(2N - n_1 - a)$.

Si $a - n_1 < 0, E[(n_2 - n_1) \mid n_1] = (a - n_1)n_1/(n_1 + a)$.

Notons $p(n_1)$ la probabilité que le premier échantillon ait la taille n_1 dans B. On a:

$$E(n_2 - n_1) = \sum_{n_1} p(n_1) E[(n_2 - n_1) \mid n_1].$$

Comme les tailles de A et B sont égales, $p(n_1) = p(2a - n_1)$, d'où:

$$E(n_2 - n_1)$$
$$= \sum_{n_1 < a} p(n_1)\{E[(n_2 - n_1) \mid n_1] + E[(n_2 - n_1) \mid (2a - n_1)]\}$$
$$= \sum_{n_1 < a} p(n_1)(a - n_1)[(N - n_1)/(2N - n_1 - a) - (2a - n_1)/(3a - n_1)]$$
$$= (2a - N) \sum_{n_1 < a} p(n_1)(a - n_1)^2/[(2N - n_1 - a)(3a - n_1)]$$
$$= (2a - N)K, K > 0.$$

Sauf dans le cas $2a - N = 0$, $E(n_2 - n_1)$ n'est pas nul et $E(n_2)/N$ est différent de a/N. La probabilité d'inclusion n'est donc pas uniforme dans $B + C$.

constitué des unités de plus petit rang selon ω_i dans chaque strate après un nombre réel indépendant des ω_i. En l'occurrence il s'agit de 0. Le retirage s'effectuait de la même manière en sélectionnant les unités de plus petit rang selon $\rho_{i,1}$, après ce nombre, dans les nouvelles strates.

Après rotation cela ne marche plus: il n'existe plus de réel indépendant des numéros tel que l'échantillon s'_1 soit constitué des unités de plus petit rang après lui. Cela est vrai même dans le cas $f_{h_1} = f_1$. Le problème est évidemment aggravé avec f_{h_1} variant par strate. L'idée qui vient alors à l'esprit est de procéder d'abord à une transformation des numéros de façon que ceux de s'_1 se retrouvent au début de $[0,1)$. On sera ensuite ramené au cas sans rotation. C'est le même genre d'idée qui est présenté par Hidiroglou, Choudhry et Lavallée (1991).

Cette transformation est assez immédiate dans le cas particulier où les mises à jour se sont faites avec la procédure A et avec les arrondis variables de la section 7.1.2. Sans retirage l'intervalle de sélection au temps t_2 aurait été:

$$\rho_{i,1} \in \left[(t_2 - t_1) f_{h_1}/D_{h_1}, (t_2 - t_1) f_{h_1}/D_{h_1} + f_{h_1}\right).$$

Le retirage se traduit par de nouvelles strates avec des probabilités f_{h_2}. Celles-ci incluent les créations d'unités entre les dates $t_2 - 1$ et t_2, auxquelles on attribue des numéros équidistants $\rho_{i,1}$, dans chaque strate h_2, indépendamment des survivants. Elles contiennent toujours les unités dont la mort est survenue depuis le tirage précédent. Il est possible de définir un nouvel échantillon s_2 de la même manière que pour le tirage de Poisson, c'est à dire par l'intervalle

$$\rho_{i,1} \in \left[a_{h_1}, a_{h_1} + f_{h_2}\right),$$

où:

$$a_{h_1} = (t_2 - t_1) f_{h_1}/D_{h_1} + \max\left[0, f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2}\right].$$

Rappelons qu'on décale de la quantité supplémentaire

$$f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2}, \quad \text{si} \quad f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2} > 0,$$

pour éviter que des unités qui viennent de sortir de l'échantillon ne s'y retrouvent trop vite.

Comme pour le tirage de Poisson, la probabilité qu'a un survivant d'être dans l'ancien et le nouvel échantillon est alors le maximum possible, à savoir:

$$\min\left(f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right), f_{h_2}\right).$$

Cependant la taille n'_{h_2} de cet échantillon est aléatoire alors qu'on veut un échantillon de taille fixée n_{h_2}. On l'obtient en sélectionnant, dans chaque nouvelle strate h_2, après avoir retranché les morts, les n_{h_2} unités de plus petit rang selon $\eta_{i,1} = \rho_{i,1} - a_{h_1}$. Ce numéro joue donc, pour le retirage, le même rôle qu'a joué ω_i au premier tirage.

Si, par contre, on a choisi la procédure A avec arrondi fixe dans la rotation ou si on a choisi la procédure B, on doit repartir du rang des unités de $\mathbf{h}_1$ lors de la dernière mise à jour. Il s'agit du rang selon ω_i avec la procédure A ou du rang selon $\rho_{t_2} - 1$ avec la procédure B. Posons $N_{\mathbf{h}_1}$ la taille de la population à la date $t_2 - 1$. Soit $R_{\mathbf{h}_1, d}$ le rang de l'unité précédent celle de plus petit rang dans s'_1 et $R_{\mathbf{h}_1}(i)$ le rang de l'unité i. Le numéro servant à classer les unités dans les nouvelles strates devient:

$$\eta_{i,1} = \frac{R_{\mathbf{h}_1}(i) - 1 - a_{\mathbf{h}_1} + \delta_{\mathbf{h}_1}}{N_{\mathbf{h}_1}} \quad \text{modulo } 1,$$

où:

$$a_{\mathbf{h}_1} = R_{\mathbf{h}_1, d} + \max\left(0, n'_{\mathbf{h}_1}/N_{\mathbf{h}_1} - f_{h_2}\right).$$

Avec la procédure A on peut garder $\delta_{\mathbf{h}_1} = \theta_{\mathbf{h}_1}$ alors qu'on fait un choix de $\delta_{\mathbf{h}_1}$ cohérent avec le dernier arrondi si la procédure B est appliquée. Mais la rotation fait que ce choix a une incidence faible sur le recouvrement et ce serait presque aussi bien de tirer au hasard dans $[0,1)$.

9. CONCLUSION

Les algorithmes basés sur les numéros équidistants ne produisent pas des TAS. Les probabilités d'inclusion du premier ordre ne sont pas exactement contrôlées et celles du second ordre sont inconnues. Lors des changements de strate, il subsiste une « trace » des anciennes strates dans les nouvelles. L'application des formules du TAS pour estimer la variance aboutit à des résultats biaisés, généralement dans le sens de la surestimation. Cependant on pense que l'amélioration des recouvrements lors des retirages, procurée par les algorithmes basés sur les numéros équidistants l'emporte sur l'inconvénient d'une estimation biaisée de la variance et des intervalles de confiance. D'après la section 5 cet avantage est d'autant plus net que la stratification est plus fine. En particulier l'usage des numéros équidistants paraît bien indiquée avec la procédure A où les strates (b, h) risquent d'être très petites pour les vagues de naissances $(b > 1)$. L'avantage des numéros équidistants est moindre avec la procédure B. Mais le fait de rendre équidistants les numéros des naissances rend moins aléatoire le nombre de survivants repris à chaque mise à jour de l'échantillon ainsi que la durée d'inclusion.

Cependant, voyons rapidement ce qui changerait dans la maintenance si on voulait conserver un TASST. A chaque étape on doit conserver la distribution indépendante et uniforme des ω_i. D'abord les phases de mises à jour des naissances et de rotation entre retirages décrites aux sections 6 et 7 s'appliquent en conservant toujours le même ω_i

allant du rang $1 + q_h n_h + \sum_{l=0}^{r_h} n_{h,l}$ au rang $(q_h + 1)n_h + \sum_{l=0}^{r_h} n_{h,l}$.
Si $D_h = D$, on peut s'imposer en plus

$$\left| \sum_{h=1}^{H} n_{h,l} - \frac{n}{D} \right| < 1, l = 1, ..., D_h.$$

La variance du taux de rotation est alors pratiquement nulle.

Toutefois, la durée d'inclusion n'est pas contrôlée quand $v_h < 1$: on a $n_h = 0$ ou $n_h = 1$. Dans le premier cas, il n'y a pas de rotation, et dans le deuxième cas, au contraire, le temps d'exclusion peut être jugé trop bref. La méthode suivante permet d'obtenir une rotation correspondant à v_h.

7.1.2 Arrondi variable

L'échantillon $s_{h,t}$ est défini à partir des numéros rendus équidistants:

$$i \in s_{h,t} \leftrightarrow \rho_{i,1} \in \left[f_h \frac{t - t_1}{D_h}, f_h \frac{t - t_1}{D_h} + f_h \right).$$

La taille de l'échantillon varie entre $I(v_h)$ et $I(v_h) + 1$ dans la strate, et elle est indépendante des tailles dans les autres strates. On retrouve ainsi ce que deviendrait la rotation de l'échantillon préconisée par Brewer, Early, et Hanif (1984) dans le cas du tirage stratifié à taille fixe et probabilité uniforme dans chaque strate.

7.2 Rotation avec mise à jour des naissances et des morts

Pour simplifier, on suppose que chaque nouvelle vague d'enquête est accompagnée de l'introduction des naissances depuis la vague précédente et d'une rotation. La méthode bifurque en deux procédures selon qu'on veut ou non respecter exactement les durées d'inclusion D_h entre deux retirages.

7.2.1 Procédure A

Les naissances sont isolées dans des strates à part, et on attend le retirage pour soustraire les morts. Dans ce cas, chaque vague de naissances est traitée exactement comme un premier tirage après avoir attribué des nombres ω_i. Le tirage se fait en stratifiant avec la même nomenclature (h), ou avec une autre plus éclatée ou plus regroupée. Pour simplifier les notations, mais sans perte de généralité, on suppose que c'est la même nomenclature. L'indice de stratification peut alors s'écrire (b,h), où b croisé avec h indique la vague des naissances avec une modalité particulière $b = 1$ correspondant aux unités déjà existantes lors du premier tirage ou retirage précédent. On est ramené aux cas de la section 7.1 dans chaque strate (b,h) et la durée d'inclusion est respectée exactement.

Le nombre de strates, donc d'arrondis, est multiplié par le nombre de vagues de naissances. La taille de l'échantillon peut devenir assez aléatoire avec des arrondis indépendants (mais moins qu'avec le tirage de Poisson). On peut donc avoir intérêt à lier, au moins partiellement, les arrondis. Par exemple, on fait un arrondissage systématique dans la dimension h pour chaque b ou l'inverse. On conserve ensuite ces arrondis et c'est la méthode 7.1.1 qui s'applique alors plutôt que la méthode 7.1.2.

7.2.2 Procédure B

Dans la procédure B, on soustrait les morts à chaque vague d'enquête. C'est le type de mise à jour présenté à la section 6. On voudrait une durée d'inclusion fixe, mais cela est rendu difficile du fait du nombre aléatoire des morts. Tout au plus peut-on essayer de contrôler une durée d'inclusion maximale DM_h. On peut souhaiter également éviter que des unités venant de sortir de l'échantillon n'y retournent à une prochaine occasion, ce qui peut arriver si la rotation est lente. L'idée est de se ramener à l'algorithme décrit à la section 6 en retranchant d'abord de $s_{h,t}$ les unités dont la durée antérieure de séjour dans $s_{h,t}$ a atteint DM_h. Elles se trouvent le plus à gauche de l'intervalle $[\rho_{h,d_t}, \rho_{h,e_t})$ et sont mélangées avec des naissances trop récentes pour avoir atteint DM_h. Mais celles-ci doivent être quand même retranchées pour que la répartition de l'échantillon selon les générations soit correcte. Pour cela, il suffit d'attribuer aux naissances une durée antérieure de séjour fictive comprise entre 1 et DM_h, juste après avoir défini l'échantillon. Par exemple, après avoir défini $s_{h,t}$, on affecte à chaque unité de $B_{h,t}$ appartenant à l'échantillon la même durée antérieure de séjour dans l'échantillon que celle de l'unité de $U_{h,t-1}$ située immédiatement à gauche. Ensuite soit R_{h,d_t} le rang le plus élevé parmi les rangs selon $\rho_{i,t}$ des unités de l'intervalle associé à $s_{h,t}$ ayant figuré DM_h fois dans l'échantillon, on écarte les premières unités de $s_{h,t}$ jusqu'au rang R_{h,d_t} compris. Enfin, on est ramené à l'algorithme décrit à la section 6 avec, pour ρ_{h,d_t}, le numéro de l'unité de rang $R_{h,t} + 1, \rho_{h,e_t}$ restant celui de l'unité qui suit celle de dernier rang dans $s_{h,t}$.

8. RETIRAGE APRÈS ROTATION

On reprend maintenant les indices de strates h_1, h_2. On définit la stratification $\mathbf{h}_1$ en fonction de la procédure utilisée pour les mises à jour des naissances. Avec la procédure A, on met les naissances dans des strates à part, c'est la stratification définie en croisant les vagues de naissances b avec la nomenclature h_1. Avec la procédure B, $\mathbf{h}_1$ est identique à h_1. Mais on conserve les notations des quantités indépendantes de b comme f_{h_1}, D_{h_1}.

Le tirage du nouvel échantillon s_2, dans une nouvelle stratification h_2 doit être fait à la période $t = t_2$.

On commence par retrancher de l'échantillon précédent (à la période $t = t_2 - 1$) les unités qui ont atteint la durée maximale d'inclusion autorisée. Il reste un échantillon s'_1 de taille $n'_{\mathbf{h}_1}$ dont on voudrait conserver le maximum d'unités dans le retirage.

Dans le cas sans rotation examiné à la section 5, il était facile de définir le retirage parce que l'échantillon s_1 était

Remarque 1. La transformation de numéros suivant indépendamment la loi uniforme en numéros équidistants a été proposée par Brewer, Early et Hanif (1984) comme un moyen d'effectuer la rotation d'échantillons de la même manière que le tirage de Poisson avec l'avantage d'une variance plus faible de la taille de l'échantillon. Mais cette transformation est faite en prenant l'ensemble de la population, et donc ils n'ont pas abordé le problème du recouvrement maximal lors des changements de strate. Les numéros ne changent qu'à l'occasion des mises à jour des naissances et des morts, selon une procédure qui est d'ailleurs bien différente de celle qu'on propose pour les changements de strate.

Remarque 2. Dans la démonstration que l'on vient de faire, il n'est pas nécessaire que les numéros soient complètement équidistants. Il suffit que les n_{h_1} unités de s_1 et les $N_{h_1} - n_{h_1}$ unités complémentaires aient leurs nouveaux numéros respectivement dans $[0, f_{h_1})$, $[f_{h_1}, 1)$. On pourrait attribuer ces nouveaux numéros de façon qu'ils suivent indépendamment la loi uniforme dans ces intervalles.

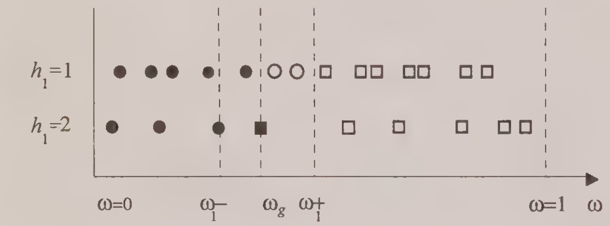


Figure 1. Recouvrement avec la méthode 1 (numéros suivant la loi uniforme).

On a représenté les unités dans g selon la valeur du nombre ω (en abscisse) et la strate h_1 du premier tirage (en ordonnée). On suppose qu'il n'y a que deux strates. Les cercles correspondent à $s_{g,1}$ et les carrés à la partie complémentaire. Les pleins correspondent à $s_{g,2}$ et les vides à la partie complémentaire. La taille de $s_{g,2}$ a été fixée à 9 ce qui définit ω_g. Dans cet exemple, on voit que deux unités ne sont pas reprises (dans $h_1 = 1$) et qu'une autre est nouvelle (dans $h_1 = 2$). La taille du recouvrement est de 8 alors que la méthode de Kish et Scott (1971) permettrait de reprendre les 9 unités dans $s_{g,1}$.

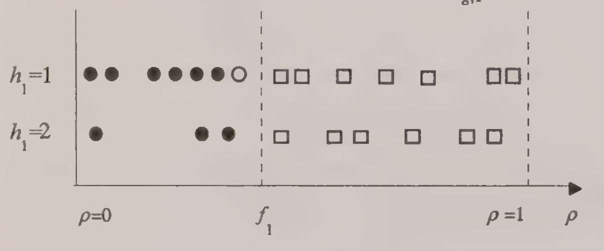


Figure 2. Recouvrement avec la méthode 2 (numéros équidistants).

On est dans la même situation que dans la figure (1), mais cette fois-ci les numéros équidistants ρ servent d'abscisses aux unités. Cette équidistance est définie dans chacune des strates h_1 entières et les trous que l'on voit apparaître dans la séquence des numéros correspondent aux unités qui ne sont pas dans g. Le premier échantillon $s_{g,1}$ est composé des unités dont ce numéro est inférieur à la probabilité d'inclusion f_1, quelle que soit la strate. Le deuxième échantillon $s_{g,2}$ est constitué des 9 unités de plus petit ρ et le recouvrement est de 9 comme pour la méthode de Kish et Scott (1971).

6. MISE À JOUR DES NAISSANCES ET DES MORTS À L'INTÉRIEUR DES STRATES

Dans cette section et la suivante on considère la stratification (h) sans référence à la période. La mise à jour des naissances et des morts à l'intérieur des strates est, dans le fond, un cas particulier de changement de strate des unités. Tout se passe comme si les naissances entraient dans les strates et que les morts en sortaient. On peut donc appliquer les méthodes précédentes. Voyons en particulier la méthode 2.

Dans une strate, la population $U_{h,t}$ d'effectif $N_{h,t}$ varie à chaque mise à jour effectuée au temps t. Notons $B_{h,t+1}$ les naissances et $D_{h,t+1}$ les morts entre t et $t+1$, on a $U_{h,t+1} = U_{h,t} + B_{h,t+1} - D_{h,t+1}$.

On considère le cas simple où les probabilités d'inclusion f_h restent uniformes dans $U_{h,t}$ et constantes. La taille $n_{h,t}$ de l'échantillon $s_{h,t}$ est un arrondi à l'entier de $N_{h,t} f_h$. Les numéros $\rho_{i,t}$ évoluent à chaque mise à jour. Juste avant la mise à jour de $s_{h,t}$, conduisant à $s_{h,t+1}$:

a) on rend équidistants les numéros $\rho_{i,t-1}$ dans $U_{h,t}$.

b) on attribue des numéros équidistants aux unités de $B_{h,t+1}$.

Soit $\rho_{i,t}$ le numéro ainsi obtenu. Une première solution consisterait à sélectionner les $n_{h,t+1}$ unités de $U_{h,t+1}$ ayant les plus petits $\rho_{i,t}$. Remarquons que ceux-ci ne sont plus équidistants parce qu'on a enlevé les morts situés au hasard.

Cependant des unités aux numéros proches de f_h peuvent sortir de l'échantillon puis y retourner à une prochaine occasion. On y remédie par un déplacement vers la droite de l'intervalle de sélection. Soient ρ_{h,d_t} le numéro de l'unité de début de l'intervalle de sélection pour $s_{h,t}$ et ρ_{h,e_t} celui de l'unité qui suit immédiatement dans $U_{h,t}$ l'unité de fin de cet intervalle. Autrement dit l'échantillon $s_{h,t}$ consiste en l'intervalle fermé à gauche et ouvert à droite $[\rho_{h,d_t}, \rho_{h,e_t})$. Entre t et $t+1$, le nombre d'unités de $U_{h,t+1}$ appartenant à cet intervalle devient $m_{h,t+1}$. Si $n_{h,t+1} \geq m_{h,t+1}$, le début de l'intervalle pour $s_{h,t+1}$ est fixé à l'unité de numéro ρ_{h,d_t}, sinon on déplace l'intervalle de façon que sa fin soit l'unité de numéro ρ_{h,e_t}. On subit donc une légère rotation involontaire.

7. ROTATION ENTRE DEUX RETIRAGES

7.1 Rotation sans mise à jour des naissances et des morts

On peut alors se donner un temps d'inclusion D_h entier et constant dans la strate. On a deux variantes selon qu'on garde le même arrondi ou qu'on le fait varier.

7.1.1 Arrondi fixe

On a donc une taille n_h strictement fixe pendant la rotation. On divise n_h en D_h nombres entiers $n_{h,l}$, ($l = 1, ..., D_h$) tels que $|n_{h,l} - n_h/D_h| < 1$. Soient q_h le quotient et r_h le reste de la division de $t - t_1$ par D_h et soit $n_{h,0} = 0$. L'échantillon au temps t comprend les unités

La méthode 1
Utilisation de numéros indépendants suivant la loi uniforme

On attribue aux unités, dès leur naissance, des nombres ω_i suivant la loi uniforme dans $[0,1)$ et indépendants, comme pour le tirage de Poisson. Le premier échantillon s_1 s'obtient en sélectionnant, par exemple, les n_{h_1} unités de plus petit rang selon ω_i dans chaque strate. Avec cet algorithme, le recouvrement maximal s'obtient également en sélectionnant les n_{h_2} unités de plus petit rang selon ω_i dans chaque strate h_2. Il est par ailleurs évident que ces deux tirages sont des TASST.

Il est aussi évident qu'on ne peut pas avoir un plus grand recouvrement avec cet algorithme. De plus, *on fait la conjecture qu'il n'est pas possible de faire mieux, pour des TASST, quel que soit l'algorithme.*

Par contre le recouvrement est plus faible en espérance qu'avec la méthode de Kish et Scott (1971), au moins dans le cas particulier où les probabilités d'inclusion du premier ordre dans s_1 sont uniformes. En effet, on n'a pas alors nécessairement dans $g s_{g,2} \subseteq s_{g,1}$ ou $s_{g,2} \supseteq s_{g,1}$, $n_{g,1,2} = n^{+}_{g,1,2}$ et *la perte de recouvrement est d'autant plus grande que les strates sont petites au premier tirage.*

Montrons-le, toujours dans le cas particulier d'une probabilité d'inclusion uniforme f_1 dans s_1. Posons ω_{h_1} la plus grande valeur de ω_i pour les unités de s_1 dans la strate h_1, et ω_g la plus grande valeur de ω_i pour les unités de s_2 dans la strate g. Soient $\omega_1^- = \min(\omega_{h_1})$ et $\omega_1^+ = \max(\omega_{h_1})$. Si $\omega_g \leq \omega_1^-$ on a $s_{g,2} \subseteq s_{g,1}$ et si $\omega_g \geq \omega_1^+$ on a $s_{g,2} \supseteq s_{g,1}$. Dans les deux cas on a bien $n_{g,1,2} = n^{+}_{g,1,2}$. Le risque de ne pas atteindre la borne n'existe que si $\omega_1^- \leq \omega_g \leq \omega_1^+$. Dans ce cas, on n'a plus nécessairement $s_{g,2} \subseteq s_{g,1}$ ou $s_{g,2} \supseteq s_{g,1}$: voir la figure (1), où on n'a considéré que 2 strates h_1. La perte de recouvrement est d'autant plus grande que la quantité $\omega_1^+ - \omega_1^-$ est plus grande en espérance, donc que les strates h_1 sont petites.

La méthode 2
Utilisation de numéros équidistants

Si on accepte de ne pas conserver un TASST, comment modifier la méthode précédente pour obtenir le même recouvrement que la méthode de Kish et Scott (1971), au moins quand on a la probabilité d'inclusion uniforme f_1 dans s_1 ? On a vu que la perte de recouvrement venait de l'écart entre les ω_{h_1}. Il suffit de transformer les ω_i en nouveaux numéros $\rho_{i,1}$ de façon que les ρ_{h_1} correspondant aux ω_{h_1} soient aussi proches que possible d'une valeur commune, soit f_1. Plus précisément, on souhaiterait avoir l'équivalence:

$$\{i \in s_1 \leftrightarrow R_{h_1}(i) \in [1, ..., n_{h_1}]\} \leftrightarrow \rho_{i,1} \in [0, f_{h_1}),$$

où $R_{h_1}(i)$ est le rang selon ω_i dans h_1 de l'unité i. Une solution est donnée par la transformation:

$$\rho_{i,1} = \frac{R_{h_1}(i) - 1 + \theta_{h_1}}{N_{h_1}} \tag{5.1}$$

où θ_{h_1} est un nombre réel vérifiant:

$$\begin{cases} \theta_{h_1} \in [0, \varphi_{h_1}), n_{h_1} = I(\nu_{h_1}) + 1, \\ \theta_{h_1} \in [\varphi_{h_1}, 1), n_{h_1} = I(\nu_{h_1}). \end{cases}$$

La transformation fait donc intervenir l'arrondi des ν_{h_1} examiné à la section 4. Le tirage de s_2 s'effectue comme celui de s_1 sauf que les $\rho_{i,1}$ jouent maintenant le rôle des ω_i: dans chaque nouvelle strate g on définit des tailles arrondies $n_{g,2}$ et on sélectionne les $n_{g,2}$ unités de plus petit rang selon $\rho_{i,1}$. Notons que ces rangs sont différents de ceux induits par ω_i.

Supposons toujours une probabilité d'inclusion uniforme dans s_1. Soit ρ_g la valeur de $\rho_{i,1}$ pour l'unité de rang $n_{g,2}$ dans g. Si $\rho_g \in [0, f_1)$, on a $s_{g,2} \subseteq s_{g,1}$. Sinon $s_{g,2} \supseteq s_{g,1}$. Dans ce cas particulier, on atteint donc le recouvrement maximal $n^{+}_{g,1,2}$ comme dans la méthode de Kish et Scott (1971) et contrairement à la méthode 1. On illustre par les figures 1 et 2 comment la transformation en numéros équidistants permet d'augmenter le recouvrement par rapport à la méthode 1.

On applique le même algorithme quand les probabilités d'inclusion dans s_1 ne sont pas uniformes. Contrairement à la méthode de Kish et Scott (1971), on n'a pas besoin de fixer la taille du nouvel échantillon à l'intérieur des sous-ensembles où ces probabilités sont uniformes. C'est un autre avantage et on pense que cela augmente le recouvrement.

Malgré tout, le recouvrement obtenu par cet algorithme reste inférieur, en espérance, à celui d'un tirage de Poisson qui aurait les mêmes probabilités d'inclusion. Pour avoir, en espérance, le même recouvrement qu'avec le tirage de Poisson il suffirait de définir $s_{g,2}$ par $\rho_{i,1} \in [0, f_g)$. En effet on aurait alors $\Pr(i \in s_1 \cap s_2) = \min(f_{h_1}, f_g)$, mais le tirage ainsi obtenu ne serait plus de taille fixe.

Les retirages suivants, après de nouvelles mises à jour, se font en itérant le procédé. Par exemple, avant de tirer s_3 on calcule des numéros équidistants $\rho_{i,2}$ à partir de $\rho_{i,1}$ (et non ω_i) dans chaque strate h_2.

Le plan de sondage qui en résulte dans les nouvelles strates n'est plus un TAS. En particulier les probabilités d'inclusion des couples d'unités varient généralement en fonction des anciennes strates. Dit de façon imagée, le retirage garde « trace » de la stratification du premier tirage. Par ailleurs, les probabilités d'inclusion des unités dans $s_{g,2}$ ne valent exactement f_g, que pour l'échantillon défini par $\rho_{i,1} \in [0, f_g)$. Pour l'échantillon de taille fixe $n_{g,2}$ cette probabilité varie en fonction des tailles des anciennes strates. Comme dans la méthode de Kish et Scott (1971) on ne contrôle pas strictement ces probabilités. Mais l'écart entre f_g et la probabilité vraie devient négligeable quand $n_{g,2}$ est assez grand.

contrôler la durée d'inclusion dans la rotation, comme pour le tirage de Poisson, et d'approcher le même taux de recouvrement lors du retirage. On commencera par le problème du recouvrement lors du retirage à la section 5. Mais auparavant, il est utile de préciser certaines notions sur l'arrondissage des tailles d'échantillons par strate.

4. ARRONDISSAGE DES TAILLES D'ÉCHANTILLON PAR STRATE

Ce problème est relié aux formules d'estimation. Celles-ci utilisent les probabilités d'inclusion du premier ordre, que ce soit dans l'estimateur sans biais de Horvitz-Thompson ou dans des estimateurs calés. Soit f_h la probabilité d'inclusion par strate, et soit $v_h = N_h f_h$. Il faut un nombre entier n_h par strate. Pour cela une première méthode consiste à restreindre le choix des f_h de façon que v_h soit entier. Dans chaque strate où l'on aurait eu $v_h < 1$ on doit prendre $v_h = 1$ pour que $f_h > 0$. Mais si la stratification est très fine vis-à-vis de la taille de l'échantillon, cela se produit dans de nombreuses strates. Cela oblige, soit à augmenter la taille de l'échantillon, soit à diminuer le taux de sondage dans les autres strates, au détriment de l'efficacité.

On va utiliser une deuxième méthode, qui consiste à lier de façon plus lâche la probabilité f_h à n_h. On applique un processus d'arrondi tel que $E(n_h) = v_h$, où v_h n'est plus nécessairement entier.

Posons $I(.)$ la fonction partie entière. On doit avoir

$$\Pr[n_h = I(v_h) + 1] = \varphi_h,$$

$$\Pr[n_h = I(v_h)] = 1 - \varphi_h,$$

où $\varphi_h = v_h - I(v_h)$.

Il n'est plus alors nécessaire que $n_h > 0$ pour que $f_h > 0$. Notons que la première méthode peut être considérée comme un cas particulier de la seconde. Ces arrondis peuvent se faire de façon indépendants par strate, de façon liée par arrondissage systématique ou par la méthode de Cox (1987). Nous décrivons seulement *l'arrondissage systématique.*

Ordonnons d'abord les strates, et indiçons-les par leur rang. Soient $c_0 = 0$ et $c_h = \sum_{j=1}^{h} \varphi_j$; on tire un nombre θ dans l'intervalle $[0, 1)$, selon la loi uniforme et on prend $n_h = I(v_h) + 1$ dans les strates telles que $c_{h-1} \leq m - 1 + \theta < c_h$ pour m entier.

Ceci implique que

$$|(n_{j_1} + \ldots + n_{j_2}) - (v_{j_1} + \ldots + v_{j_2})| < 1,$$

pour tout j_1, j_2 tels que $1 \leq j_1 \leq j_2 \leq H$.

En particulier la taille globale diffère de moins d'une unité de son espérance. Ce n'est évidemment pas le cas avec des arrondis indépendants.

5. ALGORITHMES POUR LE RECOUVREMENT MAXIMAL D'ÉCHANTILLONS DE TAILLE FIXE

Les algorithmes de maintenance que nous proposons sont basés sur l'attribution de numéros équidistants. Cela n'est pas nécessaire au premier tirage, ni dans la rotation, mais est utilisé pour maximiser le recouvrement lors des mises à jour de la stratification. C'est pourquoi on examine en premier cette phase de la maintenance.

Précisons d'abord les notations et faisons quelques constatations utiles.

On tire un premier échantillon s_1 stratifié selon un critère h_1. Au bout d'un certain temps on tire un nouvel échantillon s_2 avec une stratification h_2 mise à jour. Les probabilités d'inclusion du premier ordre sont respectivement f_{h_1}, f_{h_2} et les tailles des échantillons requises par strate sont respectivement n_{h_1}, n_{h_2}. Il suffit de considérer ce qui se passe dans une nouvelle strate quelconque $h_2 = g$. Soit $s_{g,1}$ la partie du premier échantillon s_1 dans cette nouvelle strate, dont la taille $n_{g,1}$ est généralement aléatoire. Soit $s_{g,2}$ la partie du second échantillon s_2 dans cette nouvelle strate dont la taille est fixe à l'arrondi près. La taille $n_{g,1,2}$ du recouvrement ne peut dépasser la borne $n^{+}_{g,1,2} = \min(n_{g,1}, n_{g,2})$. On peut espérer trouver un procédé de retirage à probabilité d'inclusion du premier ordre dans $s_{g,2}$ uniforme permettant d'atteindre cette borne, au moins quand les probabilités d'inclusion du premier ordre dans $s_{g,1}$ sont elles aussi égales à une seule valeur $f_{h_1} = f_1$. Remarquons que, même si cette borne est atteinte, les contraintes de taille fixe diminuent le recouvrement. Cet effet est d'autant plus marqué que la stratification est fine. En effet plus l'effectif de la strate g est petit, plus le coefficient de variation de $n_{g,1}$ risque d'être grand ainsi que la proportion d'unités non reprises dans le cas $n_{g,1} > n_{g,2}$.

Il y a une manière évidente d'atteindre la borne $n^{+}_{g,1,2}$. Supposons d'abord que les probabilités d'inclusion du premier ordre dans $s_{g,1}$ soient uniformes. Si $n_{g,1} < n_{g,2}$ on ajoute $n_{g,2} - n_{g,1}$ unités à $s_{g,1}$ tirées au hasard dans le complément de $s_{g,1}$. Si $n_{g,1} > n_{g,2}$ on retranche $n_{g,2} - n_{g,1}$ unités à $s_{g,1}$ tirées au hasard. Par construction on a $s_{g,2} \subseteq s_{g,1}$ ou $s_{g,2} \supseteq s_{g,1}$, et $n_{g,1,2} = n^{+}_{g,1,2}$. Si les probabilités d'inclusion du premier ordre dans $s_{1,g}$ ne sont pas uniformes, on applique la même méthode à l'intérieur de sous ensembles où ces probabilités sont uniformes. C'est la méthode proposée par Kish et Scott (1971) à la page 468 de leur article. Ils ne précisent pas la manière de tirer au hasard mais on suppose qu'il s'agit de TAS.

Comme le signalent Kish et Scott (1971), les probabilités d'inclusion du second ordre ne sont pas uniformes et si le premier tirage est un TASST, le second tirage ne l'est plus. La probabilité d'inclusion du premier ordre, elle-même, n'est pas strictement uniforme quand g regroupe des morceaux de strates du précédent tirage: voir un exemple en annexe. Or il existe une autre méthode qui vérifie cette condition. Elle est bien connue des offices statistiques qui pratiquent la coordination d'échantillons. Par commodité on l'appelle «méthode 1».

On effectue une partition de la population en strates $U_h, h = 1, ..., H$ de tailles N_h. Dans cet article, on appellera tirage stratifié de taille fixe un ensemble de H tirages indépendants de taille fixe n_h dans chaque strate et on se limitera à des tirages à probabilité d'inclusion du premier ordre uniforme dans chaque strate. On utilisera alors la notation $f_h = \pi_i$. On appellera tirage aléatoire simple stratifié (TASST) un tirage stratifié de taille fixe avec des tirages aléatoires simples dans chaque strate.

On appelle durée d'inclusion d'une unité le nombre d'enquêtes consécutives où elle figure dans le panel. On la notera D_i, ou D_h dans le cas particulier où elle est la même pour toutes les unités d'une strate h. Quand $\pi_i \geq 0{,}5$, cette durée ne peut pas être inférieure à $\pi_i/(1-\pi_i)$. Par exemple, si $\pi_i = 0{,}7$, la durée d'inclusion est d'au moins 3. En pratique on ne ferait pas subir de rotation aux unités dont le π_i dépasse un certain seuil.

Les variables précédentes sont en plus indicées par la vague d'enquête t. La population U_t de taille N_t et l'échantillon s_t de taille n_t varient à cause des naissances et des morts, et l'échantillon varie aussi par la rotation qu'on s'impose. D'autre part, on va considérer les échantillons aux époques particulières $t = t_1$ du premier tirage et $t = t_2$ du premier retirage. Pour alléger, ils seront notés s_1, s_2 au lieu de s_{t_1}, s_{t_2}. Les algorithmes décrits pour le couple (s_1, s_2) seront valables pour les couples suivants de retirage.

3. LA SOLUTION PAR LE TIRAGE DE POISSON

Il est éclairant d'examiner comment on peut observer le schéma de maintenance du panel par tirage de Poisson. C'est le modèle dont on va chercher à se rapprocher afin de choisir une méthode de sélection.

On attribue à chaque unité i, dès sa naissance, un numéro qui est un nombre aléatoire ω_i tiré selon la loi uniforme dans $[0,1)$. Il est sous entendu dans les formules où apparaissent ces nombres que les résultats des opérations sont *modulo* 1.

Au premier tirage, à la date $t = t_1$, on sélectionne les unités telles que ω_i appartienne à l'intervalle $[0, \pi_{i,1})$ où $\pi_{i,1}$ sont les probabilités d'inclusion que l'on se donne. En l'absence de rotation, on conserve cet intervalle aux dates suivantes jusqu'au retirage. Les naissances ainsi que les morts se répartissent au hasard dans cet intervalle. Le retirage, à la date $t = t_2$ se fait en sélectionnant les unités de l'intervalle $[0, \pi_{i,2})$ où $\pi_{i,2}$ sont de nouvelles probabilités d'inclusion. La probabilité d'inclusion conjointe est égale à la longueur de l'intervalle commun, soit $\min(\pi_{i,1}, \pi_{i,2})$ ce qui est le maximum théoriquement possible d'après la formule (2.1). L'espérance du recouvrement est donc elle-même maximale.

Considérons maintenant une rotation entre le tirage et le retirage. On maintient la probabilité $\pi_{i,1}$ et on peut se fixer une durée d'inclusion $D_{i,1}$, variable selon les unités, mais fixe jusqu'au retirage. Cette contrainte est réalisée en définissant l'échantillon à la date $t(t_1 < t < t_2)$ par l'intervalle

$$\omega_i \in [(t-t_1)\pi_{i,1}/D_{i,1}, (t-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}).$$

Le taux de rotation est une variable aléatoire. Son espérance résulte des $D_{i,1}$. Elle est égale, pour un sous ensemble quelconque V de la population, à $\sum_{i\in V}(\pi_{i,1}/D_{i,1})/\sum_{i\in V}\pi_{i,1}$.

Au premier retirage à la date $t = t_2$, on pourrait définir l'échantillon par

$$\omega_i \in [(t_2-t_1)\pi_{i,1}/D_{i,1}, (t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}).$$

Toutefois, on tombe sur une difficulté pour les unités telles que

$$\pi_{i,2} < \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right),$$

et si ω_i appartient à l'intervalle

$$\left[(t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}, (t_2-t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right)\right).$$

Ces unités qui étaient précédemment dans l'échantillon le quittent, mais vont s'y retrouver à une prochaine rotation. Si on veut l'éviter, il faut faire coïncider l'extrémité du nouvel intervalle avec celui de l'ancien, et l'échantillon à la date $t = t_2$ est finalement défini par:

$$\omega_i \in [a_{i,1}, a_{i,1} + \pi_{i,2}),$$

où:

$$a_{i,1} = (t_2-t_1)\pi_{i,1}/D_{i,1} + \max\left[0, \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right) - \pi_{i,2}\right].$$

La probabilité d'inclusion conjointe est égale à la longueur de l'intervalle en commun, soit

$$\min\left(\pi_{i1}\left(1 - \frac{1}{D_{i,1}}\right), \pi_{i,2}\right).$$

C'est aussi le maximum compatible avec la rotation.

Si on poursuit la rotation avec des durées d'inclusion $D_{i,2}$ l'intervalle à la date $t > t_2$ est:

$$\left[a_{i,1} + (t-t_2)\pi_{i,2}/D_{i,2}, a_{i,1} + (t-t_2)\pi_{i,2}/D_{i,2} + \pi_{i,2}\right).$$

Le tirage de Poisson contrôle exactement la durée d'inclusion et maximise, en espérance, le recouvrement lors du retirage mais avec l'inconvénient d'une taille d'échantillon aléatoire, dans n'importe quelle sous-population. Dans ce qui suit, on recherche des algorithmes proches de ceux qui viennent d'être décrits pour le tirage de Poisson afin de les appliquer à des tirages stratifiés à tailles fixes. On essaie de

stratification et des probabilités beaucoup plus faible que la fréquence d'interrogation. Cela est généralement le cas pour des enquêtes à périodicité infra-annuelle. La vitesse des mouvements démographiques n'est pas jugée assez grande pour qu'il soit opportun de retirer l'échantillon à chaque occasion. La rotation se fait sans changement des probabilités d'inclusion et des strates entre deux retirages et elle est étalée régulièrement dans le temps pour garder une certaine continuité à la qualité des estimateurs d'évolution. Cela correspond aussi à une durée d'inclusion dont l'espérance est constante. Dans certains algorithmes, on pourra se fixer une durée constante entre deux retirages; sinon on pourra la borner supérieurement. La vitesse de rotation traduit un compromis entre l'efficacité des estimateurs d'évolution, d'autant plus grande que le taux de renouvellement est faible, et le souci de ne pas garder une unité trop longtemps dans le panel. Notons que la recherche d'un recouvrement maximal au retirage garde un sens avec la rotation: on retranche d'abord la fraction à renouveler comme s'il n'y avait pas retirage, puis on cherche le recouvrement maximal avec la partie résiduelle.

Nous examinerons plusieurs méthodes de maintenance de panel en privilégiant la maximisation du recouvrement des échantillons lors des retirages. Nous distinguerons plus particulièrement un procédé qui assigne des numéros équidistants aux unités avant chaque changement de strate.

L'article est divisé comme suit:

Après avoir rappelé des définitions et posé quelques notations à la section 2, on indique brièvement à la section 3 comment le tirage de Poisson permet de réaliser simplement et parfaitement le schéma précédent de maintenance. Ce tirage a l'inconvénient d'être de taille aléatoire, mais il sert de référence pour les tirages stratifiés de taille fixe que l'on considère ensuite.

Le plus souvent, dans ces tirages, on se fixe au départ des probabilités d'inclusion et on procède à un arrondi pour déterminer une taille entière de l'échantillon dans chaque strate. Ce problème, traité à la section 4, n'est pas négligeable quand les strates sont petites, ce qui peut arriver pour des strates de naissances. De plus l'arrondi intervient dans la méthode qu'on propose pour maximiser le recouvrement après retirage.

La section 5 traite du recouvrement maximal d'échantillons de taille fixe. On rappelle d'abord deux méthodes connues: celle de Kish et Scott (1971) et une autre basée sur l'attribution de nombres permanents indépendants suivant la loi uniforme à chaque unité. La méthode de Kish et Scott (1971) ne paraît guère adaptée à une rotation intermédiaire entre retirages. L'autre méthode qui reproduit des tirages aléatoires simples dans chaque strate n'a pas cet inconvénient, mais le recouvrement est plus faible qu'avec la méthode de Kish et Scott (1971). Finalement on propose que les numéros soient équidistants avant retirage. On obtient alors le même recouvrement qu'avec la méthode de Kish et Scott (1971) au moins dans le cas de la répartition proportionnelle, tout en facilitant les rotations intermédiaires. Cependant le recouvrement reste inférieur au recouvrement théorique maximum que l'on obtient, par exemple, avec le tirage de Poisson.

Dans les sections 6 et 7 on présente les phases intermédiaires de mise à jour des naissances et des morts et de rotation.

Pour en terminer avec la maintenance, on montre à la section 8 comment le retirage peut s'insérer entre deux phases de rotation. On présente un type d'algorithme qui prolonge après retirage la rotation avant retirage. Il est basé sur des transformations des numéros aléatoires servant aux tirages, de façon à se ramener au retirage sans rotation. Ces transformations sont particulièrement simples quand elles portent sur les numéros équidistants, mais peuvent aussi se faire avec les numéros uniformes de départ si on veut continuer avec des tirages aléatoires simples.

2. RAPPELS, DÉFINITIONS ET NOTATIONS

Soit une population, ou ensemble fini d'unités $i \in U = \{1, ..., N\}$ où N est la taille de la population.

On ne considère que des échantillons sans remise. Un échantillon est alors simplement un sous-ensemble s de U. On appelle taille de l'échantillon le nombre n d'unités qu'il contient.

Un plan de sondage ou tirage est une probabilité discrète $p(s)$ sur l'ensemble des échantillons.

On peut généraliser à des tirages conjoints de plusieurs échantillons. En se limitant à deux échantillons s_1, s_2, le tirage conjoint est la probabilité $p(s_1, s_2)$ sur l'ensemble des couples (s_1, s_2).

La probabilité d'inclusion du premier ordre d'un individu i est définie par:

$$\pi_i = \sum_{s \ni i} p(s).$$

$E(.)$ étant l'espérance eu égard au sondage, on a:

$$E(n) = \sum_{i \in U} \pi_i.$$

Dans le cas de deux échantillons avec les probabilités d'inclusion du premier ordre $\pi_{i,1}, \pi_{i,2}$, on peut définir la probabilité d'inclusion conjointe:

$$\pi_{i,1,2} = \sum_{s_1 \ni i,\, s_2 \ni i} p(s_1, s_2).$$

On a la contrainte:

$$\pi_{i,1,2} \le \min(\pi_{i,1}, \pi_{i,2}). \tag{2.1}$$

Si $i \in s_1$, la probabilité de reprise dans s_2 est $\pi_{i,1,2}/\pi_{i,1} \le \min(1, \pi_{i,2}/\pi_{i,1})$.

Dans le tirage de Poisson, les tirages des unités sont indépendants et la taille de l'échantillon est aléatoire. Sauf à la section 3, on va plutôt considérer des tirages dont la taille est fixe à un arrondi près.

Le tirage aléatoire simple (TAS) est un tirage de taille fixe où les échantillons sont équiprobables. Cela entraîne $\pi_i = n/N$.

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 117-126
Statistique Canada

Tirage et maintenance d'un panel stratifié de taille fixe

F. COTTON et C. HESSE[1]

RÉSUMÉ

Les offices statistiques constituent souvent leurs panels d'entreprises par tirages de Poisson, ou par tirages stratifiés de taille fixe et à probabilités uniformes dans chaque strate. A ces tirages correspondent des algorithmes utilisant des numéros permanents suivant une loi uniforme. Comme les caractéristiques des unités évoluent, il est nécessaire d'effectuer périodiquement des retirages tout en cherchant à conserver le maximum d'unités. La solution par tirage de Poisson est la plus simple et donne le recouvrement théorique maximal, mais avec l'inconvénient d'une taille aléatoire de l'échantillon. Par contre, dans le cas du tirage stratifié de taille fixe, les changements de strates occasionnent des difficultés venant justement de ces contraintes de taille fixe. Une première difficulté est qu'on diminue le recouvrement, d'autant plus que la stratification est fine. Or c'est ce qui risque de se produire si les naissances constituent des strates à part. On montre comment le fait de rendre équidistants les numéros avant les retirages peut servir à corriger cet effet. L'inconvénient, assez faible, est que dans chaque strate le tirage n'est plus un tirage aléatoire simple ce qui rend moins rigoureuse l'estimation de la variance. Une autre difficulté est de concilier le retirage avec une rotation éventuelle des unités dans l'échantillon. On présente un type d'algorithme qui prolonge après retirage la rotation avant retirage. Il est basé sur des transformations des numéros aléatoires servant aux tirages, de façon à se ramener au retirage sans rotation. Ces transformations sont particulièrement simples quand elles portent sur les numéros équidistants, mais peuvent aussi se faire avec les numéros suivant une loi uniforme.

MOTS CLÉS: Panel; tirage stratifié de taille fixe; tirage aléatoire simple stratifié; recouvrement maximal; rotation de l'échantillon; numéros équidistants.

1. INTRODUCTION

On considère les tirages successifs d'échantillons destinés à suivre dans le temps l'évolution de sommes de variables, plus généralement de fonctions de sommes, dans une population. Par exemple, il s'agit d'une population d'entreprises ou d'établissements dont on veut suivre l'évolution mensuelle des ventes. L'idéal serait de pouvoir conserver un échantillon constant, mais des mouvements démographiques l'empêchent et on peut ne pas le souhaiter, compte tenu de la charge que supportent les enquêtés.

Les méthodes de sélection des unités présentées dans cet article sont soumises aux trois contraintes suivantes.

Premièrement, il est nécessaire d'introduire régulièrement les naissances et de tenir compte des morts.

Deuxièmement, le tirage fait intervenir des caractéristiques évolutives d'unités, comme la taille ou l'activité principale d'entreprises. Ces caractéristiques peuvent servir à moduler les probabilités d'inclusion. Notamment, il est souvent judicieux de faire croître ces probabilités avec la taille des unités si l'on estime des sommes de variables corrélées avec cette taille. De plus, ces caractéristiques peuvent intervenir comme critères éventuels de stratification. Dans cet article, une *strate* signifiera un sous-ensemble de la population à l'intérieur duquel le tirage est à *taille fixe*, *à un arrondi près*. Or les critères ayant servi à la stratification du premier tirage deviennent «inexacts» comme l'activité principale de l'unité, ou de moins en moins corrélés avec les variables d'intérêt comme la taille. Il s'ensuit une augmentation progressive de la variance des estimations. Pour y remédier, il convient de faire de temps en temps un *retirage* de l'échantillon après avoir mis à jour la stratification et calculé de nouvelles probabilités d'inclusion. Ceci doit être fait en essayant de conserver un maximum d'unités. Mais, fatalement, des unités seront écartées et d'autres seront introduites, principalement à cause des changements de probabilités d'inclusion. Mais cela arriverait aussi du fait des changements de strates, même si les probabilités d'inclusion restaient constantes.

Troisièmement, on souhaite répartir les charges d'enquêtes sur un plus grand nombre d'unités. On se fixe une durée limite d'inclusion dans le panel. Au-delà l'unité est remplacée par une autre choisie parmi celles qui n'y ont jamais été, ou qui sont les plus anciennes à en être sorties. On appelle *rotation* cette évolution de l'échantillon. Elle est généralement lente et régulière. Les différentes méthodes pour effectuer cette rotation sont bien connues dans les offices statistiques. Elles consistent principalement à attribuer, dès le départ, un numéro aléatoire permanent à chaque unité de la population. Les échantillons successifs sont définis par des intervalles sur ces numéros ou sur les rangs induits par ces numéros.

On appelle «*panel*» la suite chronologique des échantillons résultant de ces opérations de mise à jour, et *maintenance* du panel l'ensemble des opérations de mise à jour.

Le schéma de maintenance présenté dans cet article est analogue à celui de Hidiroglou, Choudhry et Lavallée (1991). Il correspond à une fréquence de mise à jour de la

[1] F. Cotton, Institut National de la Statistique et des Études Économiques, Département de l'Informatique et C. Hesse, Institut National de la Statistique et des Études Économiques, Département «Système Statistique d'Entreprises», 18 boulevard Adolphe-Pinard, 75675, Paris, Cedex 14.

HOAGLIN, D.C., MOSTELLER, F., et TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley.

HOGG, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of The American Statistical Association,* 69, 909-923.

HOGG, R.V. (1982). On adaptive statistical inferences. *Communication in Statistics*, 11, 2531-2542.

HOGG, R.V., BRIL, G.K., HAN, S.M., et YUL, L. (1988). An argument for adaptive robust estimation. *Probability and Statistics: Essays in Honor of Franklin A. Graybill.* Amsterdam: North-Holland/Elsevier, 135-148.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1995). Outliers in business surveys. Dans *Business Survey Methods*. New York: John Wiley.

MOBERG, T.F., RAMBERG, J.S., et RANDLES, R.H. (1980). An adaptive multiple regression procedure based on M-estimators. *Technometrics*, 22, 213-224.

RAVALET, P. (1996). L'estimation du taux d'évolution de l'investissement dans l'enquête de conjoncture: analyse et voie d'amélioration. Document de travail de l'INSEE Méthodologie Statistique, 9604.

RIVEST, L.P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de régression sont estimés simultanément. *La Revue Canadienne de Statistique*, 17, 141-153.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SOHRE, P. (1995). The Adaptive KOF Procedure for the Estimation of Industry Investment. 22nd CIRET Conference, Singapore.

WELSH, A.H., et RONCHETTI, E. (1994). Bias-Calibrated Estimations of Totals and Quantiles from Sample Surveys Containing Outliers. Rapport Technique, Department of Econometrics, University of Geneva, Switzerland.

asymétrie vers la droite de la distribution des résidus. Par ailleurs, ces nouvelles estimations se rapprochent plus de celles de l'E.A.E que des Comptes Nationaux. Ceci n'est guère très surprenant vu l'excellente corrélation entre les données individuelles de l'E.A.E et les réponses obtenues à l'enquête. Les écarts en 1991 et 1994 par rapport aux comptes demeurent pour l'instant inexplicables. En dehors de l'année 1994, les estimations obtenues avec la fonction de Cauchy sont tout à fait acceptables dans les secteurs des biens intermédiaires, de l'automobile, et dans une moindre mesure des biens d'équipement professionnel. En revanche, dans les biens de consommation, les résultats sont assez éloignés des Comptes Nationaux. On se heurte ici vraisemblablement à un problème de qualité de l'échantillon. Ce secteur est très hétérogène et quelques activités comme l'imprimerie sont mal couvertes par l'enquête.

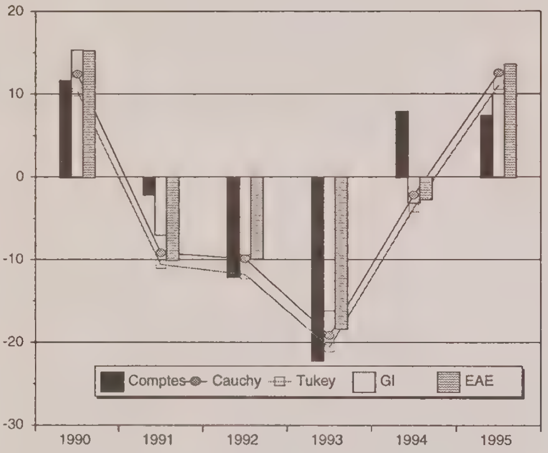


Figure 5. Taux de croissance de l'investissement en valeur dans l'industrie manufacturière

6. CONCLUSIONS

Cet article présente une justification théorique de la procédure actuellement utilisée pour dépouiller l'enquête Investissement, et notamment du principe d'exclusion des points extrêmes ou grands investisseurs. Toutefois la stratégie de repondération de l'estimateur linéaire à la Hidiroglou et Srinath (1981) présente ici des insuffisances, liées pour l'essentiel à l'identification et au traitement des points extrêmes représentatifs. La dichotomie entre individus extrapolables et grands investisseurs apparaît trop radicale et conduit à un manque de robustesse, puisque la courbe d'influence de cet estimateur n'est pas continue.

En revanche, l'hypothèse d'un modèle linéaire de superpopulation et son estimation par les GM-estimateurs nous ont semblé être d'un grand intérêt méthodologique et pratique. L'insertion de ces techniques au sein d'une procédure adaptative permet, de plus, de disposer d'un estimateur robuste pour un ensemble varié de situations. Suivant les principes décrits dans la littérature, la procédure proposée ici utilise des indicateurs d'épaisseur de queue et de concentration des résidus du modèle linéaire calculés sur l'échantillon, pour décider du réglage de la fonction de poids à utiliser, les résidus étant supposés par ailleurs symétriques. Les estimations réalisées avec la fonction de Cauchy ont donné des résultats satisfaisants sur l'industrie manufacturière et valident largement celles déjà publiées. Les avantages de cette méthode par rapport à celle utilisée actuellement s'expriment pour l'essentiel en termes de coûts de mise en oeuvre et d'une plus grande maîtrise de la méthodologie employée.

La procédure adaptative a été construite indépendamment de l'enquête. Aussi l'optimalité de la classification par rapport au contenu des strates n'est pas garantie. Par ailleurs, nous n'avons pas étudié la robustesse de la règle d'affectation à une classe. Cette question est importante lorsque l'on effectue plusieurs mesures successives et l'on désire en interpréter les révisions. A l'évidence, d'autres recherches sur ces méthodes de classification sont nécessaires, pour intégrer, par exemple, l'information livrée par les estimations précédentes ou les enquêtes exhaustives sur la population étudiée.

REMERCIEMENTS

L'auteur tient à remercier Michel Hidiroglou et Dominique Ladiray pour leurs commentaires et suggestions lors de l'élaboration de cet article.

BIBLIOGRAPHIE

BREWER, K.R. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of Statistics,* 5, 93-105.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

CHAMBERS, R.L., et KOKIC P.N. (1993). Outlier robust sample survey inference. *Bulletin de l'Institut International de Statistique, actes de la 49*ième *session*, livraison 2, 55-72.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

GWET, J.P., et RIVEST, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E., ROUSSEEUW, P.J., et STAHEL, W.E. (1986). *Robust Statistics: The Approach Based on Influence Function.* New York: John Wiley.

HIDIROGLOU, M.A., et SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

La synthèse de ces résultats permet de définir les réglages à employer sur chaque classe de distribution. Ces réglages, établis pour des échantillons de taille 100 (tableau 2), restent tout à fait acceptables pour des échantillons dont la taille est comprise entre 50 et 150.

Tableau 2
Réglage des estimateurs selon la classe des distributions des résidus ($n = 100$)

Classe	Tukey	Cauchy
I	7	7
II	4,5	4
III	3	1
IV	3	1

5. APPLICATION À L'ENQUÊTE INVESTISSEMENT

5.1 Le problème de la stratification

Les strates utilisées pour l'estimateur GI sont définies par le croisement d'une activité (18 secteurs manufacturiers) et d'une tranche de taille d'entreprise (petites, moyennes et grandes). Parmi ces 54 strates, une vingtaine environ ne regroupent jamais plus de vingt observations. Cette stratification est donc trop fine pour l'utilisation correcte de la procédure adaptative qui suppose un nombre minimal d'observations.

Comme les petites entreprises se distinguent assez nettement des moyennes et des grandes, en termes de dispersion et d'épaisseur de queue des résidus, on conserve la différentiation par taille. Des secteurs doivent donc être regroupés. La méthode, utilisée par Sohre (1995), qui consiste à regrouper après la collecte des données les secteurs ayant des paramètres (ici l'évolution moyenne de l'investissement) les plus proches, n'a pas été retenue. La proximité est en effet impossible à apprécier sur de petites strates et les regroupements obtenus sont susceptibles de changer d'une enquête à l'autre, rendant les comparaisons difficiles. Nous avons préféré redéfinir 15 nouvelles strates à partir d'un niveau de nomenclature supérieur distinguant quatre secteurs seulement: biens intermédiaires, biens d'équipement professionnel, automobile et biens de consommation.

5.2 Caractéristiques des strates

L'hypothèse d'une variance des résidus indépendante de x dans le modèle ξ ne peut être acceptée. Le choix de γ dans la fonction η s'effectue de façon à ce que la courbe des résidus (en valeur absolue) en fonction du régresseur, lissée par la méthode du LOESS, ne présente pas de tendance (Cleveland 1979). Pour la strate – biens intermédiaires, taille moyenne – à l'enquête d'avril 1995 (voir figure 4), $\gamma = 1{,}3$ est un compromis acceptable entre l'apparition d'une tendance à la baisse pour les x petits et l'annulation de la tendance à la hausse pour les plus grandes valeurs de x. Un examen similaire sur les autres strates a confirmé ce choix pour l'ensemble de l'industrie manufacturière.

Dans chaque strate, la distribution des résidus apparaît systématiquement à queue plus épaisse que la loi normale, sans être à queue très épaisse. Dans un même secteur d'activité, l'indice d'épaisseur de queue décroît avec la taille des entreprises. La grande majorité des strates représentant les petites et moyennes entreprises ont été affectées dans la classe 2. Les grandes entreprises présentent plus souvent des distributions de résidus à queue peu épaisse, proches soit de la loi normale (classe 1), soit de la loi double exponentielle (classe 4). La classe 2 est largement majoritaire et représente 75 % des cas. Seulement 20 % des distributions sont reconnues à queue peu épaisse et affectées en proportions égales dans les classes 1 et 4. En revanche, les distributions à queue très épaisse (classe 3) sont exceptionnelles (moins de 5 % des cas). S'il semble exister une certaine rémanence de la classification, celle-ci n'est pas parfaite. Et les changements sont bien réels puisqu'ils résistent à une légère modification des frontières entre les classes. Ceci justifie donc parfaitement l'utilisation d'une procédure adaptative.

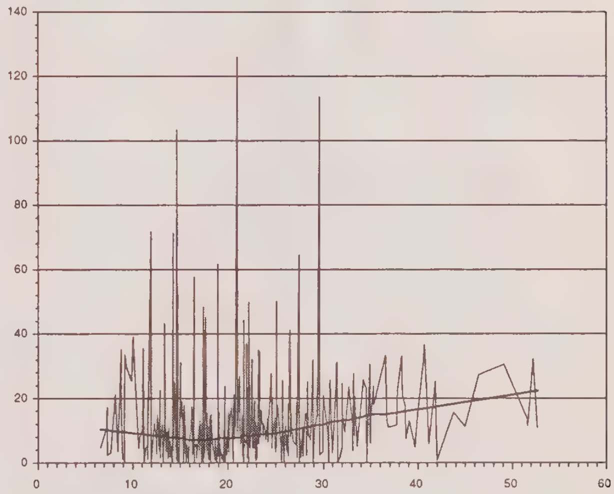


Figure 4. Valeur absolue des résidus ($\gamma = 1{,}3$, biens intermédiaires taille 2, Avril 95)

5.3 Les estimations réalisées

La procédure d'estimation basée sur (5), appliquée aux six enquêtes couvrant la période 1990-1995, a donné les résultats portés sur la figure 5. On y trouvera aussi les estimations de la Comptabilité Nationale, celles obtenues par l'estimateur GI ainsi que celles calculées issues de l'Enquête Annuelle d'Entreprise (Œ.A.E) qui est exhaustive.

Sur l'ensemble de l'industrie manufacturière, les résultats de la procédure adaptative sont comparables à ceux de l'estimateur GI. La fonction bicarrée conduit à des estimations toujours inférieures à celles obtenues avec la fonction de Cauchy. Avec un point de rejet fini, la fonction de Tukey est en effet moins influencée par la légère

l'estimateur de Chambers, de plus, confirme cette observation. Seul le cas symétrique est considéré ici; le biais des estimateurs définis par (5) est nul par conséquent.

4.3 Classification des distributions et réglage de l'estimateur

La définition de la règle de décision s'est appuyée sur l'étude de huit distributions symétriques particulières illustrant diverses situations d'épaisseur de queue et de concentration (voir tableau 1). La famille des distributions contaminées $CN(\alpha, K)$, de fonction de répartition $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/K)$ où Φ est la fonction cumulative de la loi $N(0,1)$, nous a paru intéressante car ces lois donnent une bonne représentation de données réelles (Hoaglin et coll. 1983 chap. 10) et notamment celles de l'enquête investissement (Ravalet 1996). Gaussienne en leur milieu, elles contiennent néanmoins plus d'observations extrêmes que la loi normale $N(0,1)$.

Tableau 1
Huit distributions particulières

		$\tau(,05)$	pk
1	loi normale	2,59	2,76
2	loi contaminée $CN(.05,3)$	2,94	2,83
3	loi double exponentielle	3,28	3,41
4	loi contaminée $CN(.05,10)$	4,47	2,85
5	loi contaminée $CN(.10,10)$	5,42	3,05
6	loi contaminée $CN(.20,10)$	5,64	4,44
7	loi Slash	7,65	4,19
8	loi de Cauchy	7,82	4,78

Les deux indicateurs $\tau(0{,}5)$ et pk, ont été simulés sur ces huit lois, et ce, pour plusieurs tailles d'échantillon. Le graphe de $(\tau(0{,}5), pk)$ permet de distinguer quatre groupes de distributions: les distributions à queue peu épaisse et peu concentrée du type loi normale ou $CN(,05, 3)$, les distributions à queue épaisse du type $CN(,05,10)$, $CN(,10,10)$, et $CN(,20,10)$, puis les distributions à queue très épaisse du type Slash et Cauchy, et enfin les distributions concentrées comme la loi double exponentielle. Ces quatre classes sont définies (voir figure 2) par les frontières d'équation:

Classe I: $\tau(0{,}5) \le 3{,}6 - \dfrac{14}{n}$ et $pk \le 3{,}20$

Classe II: $3{,}6 - \dfrac{14}{n} < \tau(0{,}5) \le 5{,}8 - \dfrac{35}{n}$

Classe III: $5{,}8 - \dfrac{35}{n} < \tau(0{,}5)$

Classe IV: $\tau(0{,}5) \le 3{,}6 - \dfrac{14}{n}$ et $pk > 3{,}20$.

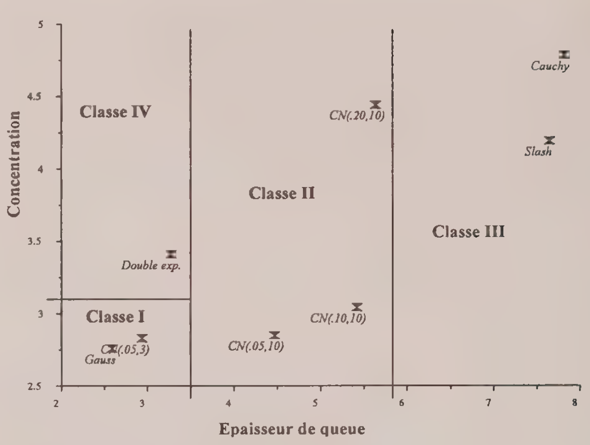


Figure 2. Quatre classes de distributions

L'ultime étape consiste à fixer le réglage des deux estimateurs dans chaque classe. Puisque l'on ne s'intéresse qu'au cas symétrique, le paramètre b de la fonction de Cauchy est nul. Par simulations, on a déterminé pour les huit lois de référence les constantes c optimales des fonctions de Tukey et de Cauchy (*i.e.*, minimisant la variance de ces estimateurs ou, ce qui revient au même ici, leur écart quadratique moyen). Celles-ci diminuent bien avec l'épaisseur de queue, si l'on excepte naturellement le cas de la loi double exponentielle qui requiert un réglage voisin de ceux utilisés pour les lois Slash et Cauchy.

L'estimateur de Tukey est plus efficace sur les lois normale ou contaminées, mais il nécessite en général un réglage plus fin. La figure 3 montre l'exemple de la loi contaminée $CN(,10, 10)$. Enfin, si le choix de la constante apparaît relativement critique pour les lois à queue épaisse ou concentrées, une large bande de valeur est envisageable pour les lois proches de la normale.

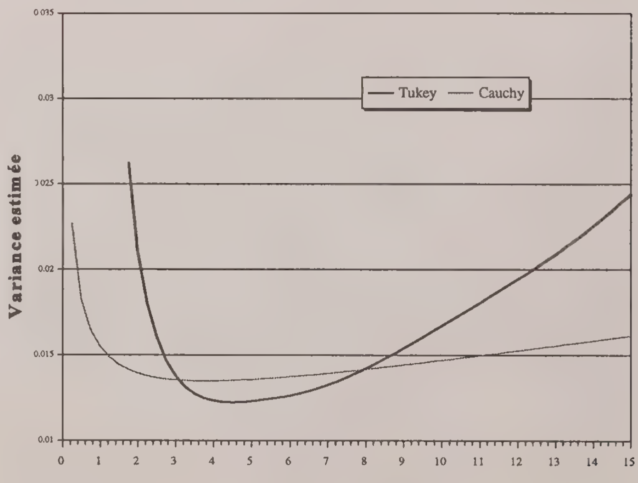


Figure 3. Variance des estimateurs de Tukey et de Cauchy pour la loi $CN(,10, 10)$ $(n = 100)$

4.1 Choix de la fonction ψ

Les fonctions monotones du type Huber n'assurant pas une protection suffisante contre les points extrêmes, seules les fonctions redescendantes ont été prises en considération. Parmi celles-ci, on a retenu les fonctions de Cauchy généralisée (utilisées notamment par Moberg et coll. 1980 pour approximer les fonctions lambda généralisées) et bicarrée de Tukey:

$$\psi_c(r) = \frac{cr}{(b+r)^2 + c}, \ \forall r$$

et

$$\psi_T(r) = \frac{r}{c}\left(1 - \frac{r^2}{c^2}\right)^2, \ \forall |r| \le c.$$

Ces deux estimateurs se différentient nettement dans le traitement des points extrêmes (voir figure 1). La fonction bicarrée suit l'identité plus longtemps que la fonction de Cauchy mais présente en revanche un point de rejet fini: les résidus au-delà de $c^*\sigma$ n'interviennent pas dans l'estimation alors que la fonction de Cauchy leur accorde une certaine représentativité. Le paramètre b permet, en principe, de contrôler l'asymétrie de ψ en fonction de celle des résidus.

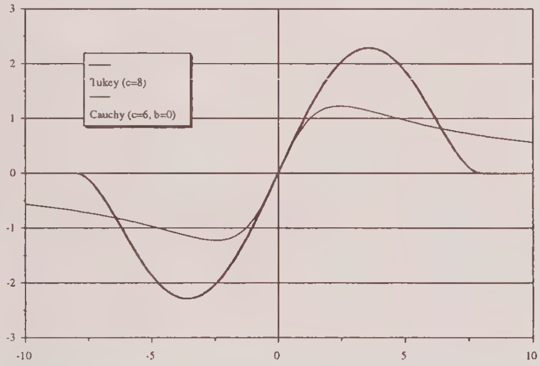


Figure 1. Fonction de Cauchy et de Tukey

4.2 Paramètre d'échelle, algorithme de calcul et critères de sélection

De façon générale un estimateur $\hat{\sigma}$ de dispersion est défini par une équation implicite $\sum\chi(r_i/\hat{\sigma}) = 0$, où χ est une fonction paire. Il s'agit donc de résoudre le système d'équations non linéaires en $(\hat{\beta}, \hat{\sigma})$ suivant:

$$\begin{cases} \sum_i \psi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0 \\ \sum_i \chi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0. \end{cases} \tag{6}$$

Rivest (1989) montre sur quelques exemples que la résolution du système (6) peut poser des difficultés en raison d'une éventuelle multiplicité des solutions, même dans le cas d'une fonction ψ monotone. Suivant ses recommandations, nous procédons en deux étapes. Dans un premier temps, le paramètre de dispersion σ est estimé à l'aide de la médiane des valeurs absolues (MAD) des résidus définis à partir de la médiane des taux d'évolution individuels. Ensuite β est calculé par (4) en utilisant la valeur de σ trouvée précédemment.

Pour la résolution de (4), nous avons préféré l'algorithme de repondération à l'algorithme de Newton-Raphson, car il semble converger plus facilement, notamment lorsque la constante de réglage est petite.

L'efficacité d'une procédure adaptative reposant sur celle du processus décisionnel, la plus grande attention doit être portée sur la nature, la qualité et la robustesse des informations commandant le choix de l'estimateur.

L'épaisseur de queue est un indicateur indispensable car elle renseigne sur l'importance relative des points extrêmes dans l'échantillon, donc dans la population (voir Hoaglin et coll. 1983, chap. 10). On a retenu comme indicateur d'épaisseur de queue la proposition de Hogg (1974):

$$\tau(p) = \frac{\bar{U}(p) - \bar{L}(p)}{\bar{U}(0,5) - \bar{L}(0,5)}$$

$\bar{U}(p)$ (resp. $\bar{L}(p)$) est la moyenne des np plus grandes (resp plus petites) statistiques d'ordre, en utilisant une interpolation linéaire lorsque np n'est pas entier. On a choisi $p = 0{,}05$; pour la loi normale $\tau(,05)$ vaut 2,59.

De plus, il nous a semblé important, comme Hogg et coll. (1988), de tester la présence éventuelle d'une distribution du type double exponentielle, en mesurant la concentration des résidus par l'indicateur pk suivant:

$$pk = \frac{\bar{X}(1-\beta, 1-\alpha) - \bar{X}(\alpha, \beta)}{\bar{X}(0{,}5,\ 1-\beta) - \bar{X}(\beta, 0{,}5)}$$

où $\bar{X}(a, b)$ est la moyenne des statistiques d'ordre entre la na-ième et la nb-ième, avec des grandeurs interpolées si na ou nb ne sont pas entiers. On a retenu $\alpha = 0{,}05$ et $\beta = 0{,}15$, soit $pk = 2{,}7$ pour une distribution normale.

Enfin, des études (Moberg et coll. 1980, Hogg et coll. 1988) ont souligné l'importance de la dissymétrie des distributions. En effet, en présence de résidus asymétriques, le biais des estimateurs robustes peut être important, rendant ainsi leur utilisation délicate (Chambers et Kokic 1993). Dans l'enquête Investissement de l'INSEE, les résidus sont théoriquement asymétriques puisque minorés ($r = y - \beta x \ge -\beta x$). Toutefois, nous avons constaté empiriquement que cette asymétrie était très légère et qu'elle pouvait être négligée sans dommages. L'échec de la correction d'un éventuel biais par la fonction ψ_E dans

$$\sum_s w\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\psi\left(\frac{r_i}{\sigma}v\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\right)\frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

avec

$$r_i = \frac{y_i - \hat{\beta}_R x_i}{\sqrt{\eta(x_i)}}.$$

Un choix habituellement retenu est la forme de Mallows: $v(t) = 1$ et $w(t) = 1/t$. Un estimateur robuste $\hat{\beta}_R$ vérifiera donc l'équation implicite

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right) = 0. \tag{4}$$

En général, le paramètre σ est inconnu et doit être remplacé dans cette expression par une estimation robuste $\hat{\sigma}$ de la dispersion des résidus

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = \sum_i \psi\left(\frac{r_i}{\hat{\sigma}}\right) = 0.$$

L'estimateur du total sera finalement:

$$\hat{Y}_{\beta R} = \sum_s y_i + \hat{\beta}_R \sum_{\bar{s}} x_i. \tag{5}$$

Cet estimateur est étudié par Gwet et Rivest (1992). En général, il n'est pas sans biais par rapport au plan de sondage. Chambers (1986) propose de corriger ce biais en introduisant dans (5) un troisième terme qui l'estime de façon robuste:

$$\hat{Y}_{\text{Chambers}} = \sum_{i \in s} y_i + \hat{\beta}_R \sum_{i \in \bar{s}} x_i +$$

$$\left(\sum_{i \in s} \frac{x_i/\hat{\sigma}\sqrt{\eta(x_i)}}{\sum_{j \in s} x_j^2/\hat{\sigma}^2\eta(x_j)} \psi_E\left(\frac{y_i - \hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right)\right)\sum_{i \in \bar{s}} x_i$$

Choisir une fonction ψ_E bornée semble un bon compromis entre biais et variance de l'estimateur. Par exemple, Welsh et Ronchetti (1994) optent pour une fonction de Huber avec une constante de réglage grande $c = 15$. Mais le réglage de ψ_E, sans information préalable sur la densité des points extrêmes, est toujours délicat.

3.2 Choix de l'estimateur

Les propriétés souhaitables des fonctions ψ sont désormais bien connues par référence au problème de l'estimation d'une tendance centrale. Celles-ci doivent être bornées, continues, équivalentes à l'identité au voisinage de zéro. On distingue habituellement les fonctions strictement monotones (Huber) des fonctions redescendantes comme la fonction bicarrée de Tukey, le sinus d'Andrew et la fonction de Hampel ou de Cauchy. Parce que leur fonction d'influence tend vers zéro, ces estimateurs seront moins sensibles à la présence de points extrêmes que la fonction de Huber. La vitesse de convergence vers zéro est une caractéristique essentielle des fonctions redescendantes. Celles, nulles à distance finie (Hampel, Tukey ou Andrew) excluent les points extrêmes de l'estimation de β alors que les autres leur accordent une faible représentativité.

Le choix et le réglage de la fonction ψ sont délicats. Ils dépendent beaucoup de la nature des données et plus précisément de la distribution des résidus (Hoaglin, Mosteller et Tukey 1983, chap. 11). Une idée, ne serait-ce qu'approximative, de l'allure de la distribution des résidus devrait permettre de mieux cibler le choix et le réglage de l'estimateur, donc de rendre l'estimation plus efficace. Cette remarque intuitive est à l'origine des procédures adaptatives, présentées notamment par Hogg (1974) et (1982). L'idée est d'apprécier la nature de la distribution des résidus, calculés à partir d'une première estimation robuste (du type norme L_1 par exemple), à l'aide d'indicateurs robustes bien choisis (épaisseur de queue, asymétrie, concentration *etc.*). La donnée de ces indicateurs permet alors de choisir, selon une règle de décision prédéfinie, l'estimateur adapté à cette situation et on résout l'équation implicite (4) en prenant comme valeur initiale la première estimation robuste de β.

Le principe d'une procédure adaptative apparaît d'autant plus séduisant qu'il systématise l'étude préalable nécessaire au choix et au réglage d'un estimateur. Celle-ci peut en effet s'avérer extrêmement coûteuse si elle doit être réalisée manuellement pour chaque strate de l'échantillon et renouvelée à chaque enquête.

4. CONSTRUCTION D'UNE PROCÉDURE ADAPTATIVE

On décrit ici la construction d'une procédure adaptative pour le calcul du taux d'évolution moyen de l'investissement à partir des données de l'enquête de conjoncture. Aussi certains choix ont-ils été effectués sachant la nature et les caractéristiques propres de ces données et ne sont pas nécessairement transposables à d'autres modèles de régression. En particulier, on a retenu, après vérification sur les données, l'hypothèse de symétrie de la distribution des résidus et exclu le cas de distributions à queue fine.

La construction d'une procédure adaptative, qui s'inspire des travaux de Moberg, Ramberg et Randles (1980), s'effectue en plusieurs étapes. On choisit la fonction (ou famille de fonctions) ψ à utiliser, puis on sélectionne l'ensemble des critères servant à qualifier la distribution des résidus. La donnée de ces critères permet la construction d'une règle de classification. Enfin, à chaque classe est associé le réglage de l'estimateur à utiliser.

La valeur optimale de λ qui minimise l'écart quadratique moyen de cet estimateur, conditionnellement ou non au nombre de valeurs extrêmes dans l'échantillon, est fonction de plusieurs paramètres de la population. Sans information a priori, le choix de λ est délicat.

Appliqué au cas de l'estimateur du ratio avec variable auxiliaire x, cela s'écrit:

$$\hat{Y}_{\text{ratio}\,\lambda} = \sum_s y_i + \sum_{\bar{s}} x_i \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} + (\lambda - 1)\left(\frac{\sum_{s_1} y_i}{\sum_{s_1} x_i} - \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i}\right)\sum_{s_1} x_i \quad (3)$$

Les deux premiers termes du second membre de (3) forment une estimation du total Y, sous l'hypothèse implicite que tous les points extrêmes sont dans l'échantillon, et le troisième est une correction tenant compte de la présence éventuelle de tels points dans la population non interrogée. Cette correction est fonction du λ retenu et de la différence des comportements moyens entre les deux types d'individus estimés sur l'échantillon.

En rapprochant (2) et (3), on s'aperçoit que l'estimateur GI est formellement équivalent au cas $\lambda = 1$. L'utilisation de $\hat{Y}_{\text{GI}}$ suppose donc implicitement que les points extrêmes ont été correctement identifiés et sont tous non représentatifs. Dans Ravalet (1996), on a montré que ces deux hypothèses étaient malheureusement rarement vérifiées dans le contexte de l'enquête Investissement.

La procédure d'identification étant manuelle et le critère retenu relativement *ad hoc* en l'absence de toute hypothèse sur la population, il n'est pas exclu que certains points extrêmes échappent à la sélection. L'utilisation du ratio sur les extrapolables pose alors le problème de la robustesse de l'estimation vis à vis du choix des grands investisseurs. En outre, tous ces points ne sont vraisemblablement pas uniques. Les points atypiques, particulièrement nombreux chez les petites et moyennes entreprises, devraient plutôt être considérés comme représentatifs. Toutefois, choisir $\lambda > 1$ introduirait immanquablement la question de la robustesse du troisième terme de (3).

Des modifications de l'estimateur $\hat{Y}_{\text{GI}}$ sont envisageables pour tenter de pallier ces défauts. La moyenne sur les extrapolables peut être par exemple remplacée par un estimateur plus robuste et seuls les points non représentatifs sont déclarés grands investisseurs. Cette technique s'inscrit dans le cadre plus général des M-estimateurs où la donnée d'un modèle facilite à la fois le repérage et le traitement des points extrêmes (Lee 1995). Il ne s'agit plus alors de procéder à une dichotomie stricte entre points extrêmes et autres points mais de définir des zones de plus ou moins grande représentativité.

3. ESTIMATION ROBUSTE PAR LES GM-ESTIMATEURS

3.1 Le modèle linéaire et les GM-estimateurs

On suppose l'existence d'un modèle linéaire ξ reliant pour l'ensemble de la population U les investissements x et y aux dates $t-1$ et t.

$$\xi : y_i = \beta x_i + \epsilon_i$$

avec

$$E(\epsilon_i) = 0$$
$$E(\epsilon_i \epsilon_j) = 0 \qquad \forall\, i \neq j.$$
$$V(\epsilon_i) = \sigma^2 \eta(x_i)$$

La pente β de la droite de régression passant par l'origine dans le modèle de superpopulation s'interprète comme le taux d'évolution Θ dans la population. La variance de y est supposée fonction croissante de x et η est en général une fonction puissance: $\eta(x_i) = x_i^{\gamma}$.

Sous le modèle, le meilleur estimateur linéaire sans biais (Brewer 1963 et Royall 1970) du total est $\hat{Y}_{mc} = \sum_s y_i + \hat{\beta}_{mc} \sum_{\bar{s}} x_i$, où $\hat{\beta}_{mc} = (\sum_s x_i y_i / \eta(x_i)) / (\sum_s x_i^2 / \eta(x_i))^{-1}$ est l'estimateur des moindres carrés.

Dans le cas particulier $\eta(x) = x$, cette expression se réduit à $\hat{\beta}_{mc} = \sum_s y_i / \sum_s x_i$, estimateur du ratio. Cet estimateur sans biais n'est efficace que sous l'hypothèse de normalité des résidus et se montre peu robuste.

Les M-estimateurs (Huber 1981) permettent de définir une version robuste des moindres carrés en substituant à la fonction carré, dans le programme de minimisation, une fonction ρ croissant moins rapidement:

$$\text{Min} \sum_s \rho\left(\frac{y_i - \beta_R x_i}{\sigma\sqrt{\eta(x_i)}}\right).$$

Le M-estimateur $\hat{\beta}_R$ est la solution de l'équation implicite:

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

où

$$\psi(t) = \frac{\partial \rho(t)}{\partial t}.$$

La fonction ψ, comme la fonction de Huber $\psi(t) = \text{Max}(-c, \text{Min}(t, c))$, dépend d'une (ou plusieurs) constante de réglage $\underline{c}$ contrôlant la part des observations qui doivent être considérées comme points extrêmes. Cet estimateur sera encore sensible à la présence de valeurs extrêmes sur la variable explicative x. On définit alors une classe plus générale d'estimateurs appelés GM-estimateurs (Hampel, Ronchetti, Rousseeuw et Stahel 1986) par l'équation implicite:

$s = \{1, ..., n\}$ de taille n, et $\bar{s} = \{n+1, ..., N\}$ désigne la population non interrogée. Chaque entreprise est interrogée sur ses dépenses d'investissement pour deux années consécutives $t-1$ et t, notées respectivement x et y.

Connaissant le montant total X des investissements de l'année $t-1$ dans la population, on peut déduire de l'estimation $\hat{Y}$ du total des investissements pour l'année t, le taux d'évolution moyen des dépenses d'équipement entre $t-1$ et t:

$$\hat{\theta} = \frac{\hat{Y} - X}{X}.$$

Pour simplifier les notations, on définit le paramètre $\Theta = 1 + \theta = Y/X$, estimé par $\hat{\Theta} = \hat{Y}/X$.

L'estimateur actuellement utilisé dans l'enquête de l'INSEE s'inspire de la méthode du ratio, avec pour information auxiliaire l'investissement réalisé en $t-1$:

$$\hat{Y}_{\text{ratio}} = \frac{X}{\sum_s x_i} \sum_s y_i.$$

Cet estimateur peut s'écrire comme un estimateur linéaire pondéré:

$$\hat{Y}_{\text{ratio}} = \sum_s w_i z_i. \tag{1}$$

Dans cette expression, $w_i = X x_i / \sum_s x_j$ est le poids de l'individu i et $z_i = y_i / x_i$ l'évolution annuelle de son investissement. Un tel estimateur sera sensible à la présence de points extrêmes à la fois sur z et w. *Un point atypique* présentera une évolution z très différente de celle des autres, tandis qu'un *point influent* aura un poids w suffisamment important pour attirer, par effet levier, le taux d'évolution moyen de la strate vers son propre taux d'évolution. Le critère décisif pour qualifier une observation de point extrême étant que le produit wz soit assez grand pour perturber l'estimation $\hat{Y}_{\text{ratio}}$, la distinction entre points atypiques et points influents est bien entendu arbitraire. Le terme générique *grands investisseurs* (ou GI en abrégé) désignera l'ensemble de ces points extrêmes tandis que le terme *extrapolables* fera référence aux autres individus de l'échantillon.

Ayant réalisé une partition a posteriori de l'échantillon $s = \{\text{GI}\} \cup \{\textit{extrapolables}\}$, on estime le total des investissements du reste de la population $\bar{s}$ à partir du comportement des seuls individus extrapolables selon la méthode du ratio:

$$\hat{Y}_{\text{GI}} = \sum_s y_i + \left(\sum_{\bar{s}} x_i\right) \frac{\sum_{\{\text{extra}\}} y_i}{\sum_{\{\text{extra}\}} x_i}. \tag{2}$$

Dans (2), le poids des extrapolables $1 + \sum_{\bar{s}} x_i / \sum_{\{\text{extra}\}} x_i$ est bien strictement plus grand que celui des grands investisseurs qui vaut 1.

2.2 Sélection des Grands Investisseurs

Les grands investisseurs sont choisis, au niveau de chaque strate, en fonction de leur influence sur l'estimation de Θ selon une procédure itérative. Pour commencer, les individus sont tous supposés extrapolables et on calcule pour chacun d'eux un indice de non prise en compte, mesurant l'impact sur $\hat{\Theta}$ de son exclusion de l'échantillon, $\text{NPEC} = (\hat{Y}^i_{\text{GI}} - \hat{Y}_{\text{GI}})/X$ où $\hat{Y}^i_{\text{GI}}$ est le total estimé sans l'individu i.

L'entreprise ayant le plus grand indice NPEC en valeur absolue est déclarée grand investisseur. On réestime alors $\hat{Y}_{\text{GI}}$ avec cette nouvelle partition de U, puis on identifie le grand investisseur suivant. La sélection s'interrompt dès que tous les individus extrapolables ont une influence sur l'estimation inférieure à un seuil donné. Cette condition est d'autant plus facilement vérifiée que le nombre et la masse des observations sont importants. Inversement, elle se révélera impossible à réaliser si le nombre d'individus est trop faible; dans ce cas, le gestionnaire d'enquête veille simplement à ce qu'aucun individu n'ait une influence beaucoup plus grande que les autres, introduisant ainsi une dose de subjectivité dans la procédure.

Par ce mécanisme itératif, les phases habituelles de détection et de traitement des points extrêmes sont réalisées de façon simultanée. La principale difficulté tient dans le fait que le statut d'un individu n'est pas une qualité intrinsèque, mais dépend de la composition de l'échantillon. Celui-ci peut changer d'une enquête à l'autre. En outre, cette procédure peut conduire dans certains cas de figure (Ravalet 1996) à exclure inutilement certains individus car, à aucun moment, le statut de grand investisseur n'est remis en question.

2.3 La stratégie de repondération de l'estimateur linéaire

L'estimateur GI suit en fait de la stratégie de repondération de l'estimateur linéaire (1) présentée par Hidiroglou et Srinath (1981) sur l'exemple de l'estimation d'un total sans information auxiliaire. Ayant réalisé a priori une partition $s = s_1 \cup s_2$ de l'échantillon distinguant les points extrêmes s_1 (en nombre n_1) des autres observations s_2, les auteurs proposent de réduire, dans $\hat{Y} = (N/n)\sum_s y_i$, le poids N/n des points extrêmes à une valeur plus faible λ en posant

$$\hat{Y}_\lambda = \lambda \sum_{s_1} y_i + \frac{N - \lambda n_1}{n - n_1} \sum_{s_2} y_i$$

soit

$$\hat{Y}_\lambda = \sum_s y_i + \frac{N-n}{n-n_1} \sum_{s_2} y_i +$$
$$n_1(\lambda - 1)\left[\frac{1}{n_1}\sum_{s_1} y_i - \frac{1}{n-n_1}\sum_{s_2} y_i\right].$$

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 107-115
Statistique Canada

Une procédure adaptative d'estimation robuste du taux d'évolution de l'investissement

PHILIPPE RAVALET[1]

RÉSUMÉ

La présence d'observations extrêmes dans les données d'enquête est un problème récurrent de la statistique appliquée auquel l'enquête de l'INSEE sur l'investissement industriel est aussi confrontée. La prévision du taux de croissance des dépenses d'équipement dans l'industrie se ramène, de ce fait, à l'estimation robuste d'un total dans une population finie. Dans une première partie, cet article analyse l'estimateur actuellement utilisé dans l'enquête Investissement. Nous montrons qu'il suit une stratégie de repondération de l'estimateur linéaire. Mais la dichotomie stricte imposée entre les points extrêmes, tous supposés non représentatifs, et les autres points n'est pas entièrement satisfaisante d'un point de vue à la fois théorique et pratique. L'adoption d'une approche modélisée et l'estimation par les GM-estimateurs, appliqués au cas d'une population finie, permet de pallier ces défauts. Nous construisons ensuite une procédure adaptative robuste qui détermine l'estimateur approprié en fonction des résidus observés sur l'échantillon lorsque ceux-ci peuvent être supposés symétriques. Enfin, cette méthode est appliquée aux données de l'enquête Investissement sur la période 1990-1995.

MOTS CLÉS: Enquêtes de conjoncture; valeurs extrêmes; estimation robuste; GM-estimateur; procédure adaptative.

1. INTRODUCTION

Depuis 1952, l'Institut National de la Statistique et des Études Économiques (INSEE) réalise une enquête sur l'investissement qui fournit des estimations prévisionnelles de l'évolution des dépenses d'équipement dans l'industrie, bien avant la publication des Comptes Nationaux et des résultats d'enquêtes exhaustives. L'estimation du taux de croissance de l'investissement s'appuie sur les déclarations d'environ 2 500 chefs d'entreprise concernant leurs dépenses et intentions de commande en biens d'équipement.

La présence quasi systématique de valeurs extrêmes dans ces données constitue une difficulté majeure. Celles-ci peuvent en effet perturber gravement l'estimation du taux de croissance moyen et conduire à des résultats inacceptables. Selon Chambers (1986), on peut distinguer deux types de points extrêmes. Les points non représentatifs correspondent soit à des erreurs de mesure, que l'on s'efforce de corriger lors de la collecte des données, soit à des individus uniques dans la population. A contrario, les points extrêmes représentatifs désignent des individus curieux mais qui ne peuvent être considérés comme exceptionnels. Il en existe certainement de semblables dans la population non interrogée et l'information qu'ils contiennent doit être intégrée dans l'estimation.

Le problème posé ici s'identifie à celui de l'estimation robuste d'un total dans une population finie avec information auxiliaire, problème auquel la théorie n'apporte pas de réponse définitive. Néanmoins diverses techniques, revues dans Lee (1995), peuvent être appliquées. La méthode d'estimation actuellement utilisée dans l'enquête Investissement suit la logique de repondération de l'estimateur linéaire selon Hidiroglou et Srinath (1981). Toutefois, l'identification et le traitement des points extrêmes ne sont pas entièrement satisfaisants. En particulier, tous les points extrêmes sont supposés non représentatifs et la dichotomie entre points «normaux» et points extrêmes rend l'estimation très sensible au choix de ces derniers.

L'introduction d'un modèle linéaire de superpopulation, qui décrit l'évolution individuelle de l'investissement, permet de mieux apprécier le caractère singulier d'une observation et de définir son niveau de représentativité. Son estimation par les GM-estimateurs constitue alors une alternative séduisante à la méthode des moindres carrés dont la propriété d'absence de biais est très coûteuse en termes de variance. Le réglage de la fonction de poids dépend a priori des caractéristiques de la population selon des critères maintenant bien décrits dans la littérature. Ces caractéristiques pouvant changer d'une strate à l'autre, mais aussi au cours du temps, l'intérêt d'une procédure adaptative est évident. A partir d'une première estimation robuste, on détermine l'allure de la distribution des résidus, puis on choisit l'estimateur à utiliser selon une règle prédéfinie. Suivant Hogg, Bril, Han et Yul (1988), on construit une procédure adaptative s'appuyant sur des indicateurs d'épaisseur de queue et de concentration estimés sur l'échantillon, l'asymétrie des résidus n'étant pas envisagée. Cette procédure est appliquée sur les données de l'enquête Investissement pour la période 1990-1995.

2. L'ESTIMATEUR DE L'ENQUÊTE INVESTISSEMENT

2.1 Principe de l'estimation

Dans une population finie $U = \{1, ..., N\}$, correspondant ici à une strate de l'enquête, on tire un échantillon

[1] Philippe Ravalet, Division des enquêtes de conjoncture, INSEE, 15 Bd G. Péri, BP 100, 92244 MALAKOFF CEDEX.

REMERCIEMENTS

Cet article est le fruit des réflexions et des travaux d'une mission, animée par les auteurs, à laquelle ont collaboré: Xavier Berne, Michel David, Michel De Bie, Sophie Destandau, Jacques Leclercq, Françoise Lemoine, Catherine Marquis, Marc Simon. La mission a bénéficié de l'aide de différents services de l'INSEE. L'Unité «Méthodes statistiques» et notamment son chef, Jean-Claude Deville, méritent tout spécialement d'être cités. Les auteurs remercient également Philippe Ravalet pour son apport théorique, ainsi que la Rédaction de *Techniques d'enquête* et les deux arbitres pour leurs commentaires constructifs.

BIBLIOGRAPHIE

DECAUDIN, G., et LABAT, J.-C. (1996). Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population. Document de travail de méthodologie statistique, n° 9601, INSEE. Paris.

DESCOURS, L. (1992). Estimation de populations locales par la méthode de la taxe d'habitation. *Actes des Journées de méthodologie statistique,* 13 et 14 mars 1991, INSEE. Paris.

GUÉGUEN, Y. (1972). Estimation de la population des villes bretonnes au 1.1.1971. *Sextant*, n° 4. INSEE. Rennes.

de GUIBERT-LANTOINE, C. (1987). Estimations de population par département en France entre deux recensements. *Population*, 6, 881-910.

HOAGLIN, D.C., MOSTELLER, F., et TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

LAURENT, L., et GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE. Rennes.

LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.

STATISTIQUE CANADA (1987). *Méthodes d'estimation de la population, Canada*. N° 91-528F au catalogue. Ottawa.

8. COMPLÉMENTS

8.1 Niveaux infradépartementaux

L'utilisation de certaines sources peut devenir hasardeuse à un niveau géographique plus fin que le département, et cela pour différentes raisons: parce que les hypothèses sur lesquelles repose la méthode deviennent fragiles, parce que les effectifs sont faibles... Les statistiques scolaires sont notamment dans ce cas.

Cependant, on ne devrait pas courir trop de risques en faisant fonctionner le système pour les zones d'emploi; plus précisément pour les croisements «département * zone d'emploi» (environ 420 zones) permettant d'assurer la cohérence avec le niveau départemental.

En effet:

- on peut accepter une certaine dégradation des performances par rapport aux estimations départementales, d'autant que ces dernières devraient être de bonne qualité;
- les données tirées des fichiers de l'impôt sur le revenu devraient être d'un apport précieux;
- l'estimation tendancielle et le calage sur les estimations de niveau géographique supérieur (départementales en l'occurrence) jouent, l'une et l'autre, un rôle de garde-fou.

Notons que rien n'interdit, bien entendu, d'utiliser le système pour produire des estimations dans d'autres zonages infradépartementaux.

Au niveau départemental, il ne semble pas utile d'adapter les paramètres (poids «a priori» et normes) à la taille de la population; en revanche, pour les niveaux infradépartementaux, cette adaptation semble indispensable. Sinon on risque d'être beaucoup trop rigoureux pour les petites zones. Il semble qu'une fonction de norme du type suivant puisse convenir:

$$NO_S = \alpha P^{\beta},$$

où NO_S est la norme de la source S, P la population de la zone et α et β deux paramètres dépendant a priori de la source S. Le paramètre β est évidemment négatif. Si β vaut $-0{,}25$, la norme double lorsque la population est divisée par 16. Il semble aussi que le type de zone intervienne: ainsi le flou serait en moyenne plus important pour une commune de 50,000 habitants que pour une zone d'emploi de même taille. Les paramètres α et β sont à définir pour chaque source infradépartementale et, le cas échéant, pour chaque type de zone.

8.2 Calendrier

Le système fonctionne d'autant mieux que le nombre de sources est plus important. Toutefois, les sources relatives à une même année sont disponibles de façon échelonnée dans le temps. Le système étant capable de fonctionner avec un nombre variable de sources, on peut élaborer, au moins au niveau départemental, plusieurs ensembles d'estimations au 1[er] janvier de l'an n: par exemple, des estimations provisoires au troisième trimestre de l'année n, à partir des premières sources disponibles, puis des estimations semi-définitives au troisième trimestre de l'année $n + 1$, assises sur davantage de sources et enfin des estimations définitives au troisième trimestre de l'année $n + 2$. Différents éléments sont à prendre en compte: la lourdeur d'une campagne, l'ampleur des modifications dues à l'ajout d'une source, ampleur qui pourra être appréciée par des simulations sur les premières années de mise en œuvre du système.

8.3 Intégration d'une source supplémentaire

Le système est souple et modulaire. L'intégration d'une nouvelle source ne pose donc pas de problème particulier. Il suffit de définir la méthode permettant d'en tirer une bonne estimation du taux de solde migratoire de chaque zone. La panoplie des méthodes envisagées par la mission est assez fournie pour que, dans la plupart des cas, on puisse y trouver un type de méthode adapté à la source.

Pour déterminer les paramètres (poids «a priori» et norme) à lui attribuer dans la synthèse, on suggère de faire fonctionner le système «à blanc» avec des paramètres fixés arbitrairement, mais de façon raisonnable; il est évidemment prudent de démarrer avec une norme plutôt forte et un poids plutôt faible. L'analyse des écarts obtenus entre les taux de solde migratoire issus de cette source et les taux synthétiques permet de déterminer une meilleure norme. On peut alors adapter le poids en conséquence, en se servant, faute de mieux, d'une relation supposée de quasi-proportionnalité entre le poids et l'inverse du carré de la norme. On peut évidemment itérer ce processus, en modifiant également, le cas échéant, les paramètres des autres sources. Toutefois, les tests réalisés au niveau départemental sur la période 1982-1990 semblent montrer que les performances globales du système sont assez peu sensibles à des variations, même assez importantes, des poids «a priori»; il n'est donc pas nécessaire de déterminer ces poids avec une grande précision, ce qu'on ne pourra pas faire, de toute façon, avant le prochain recensement.

9. CONCLUSION

Le système d'estimation de population «multi-sources» présenté ici est robuste et souple, sans être trop complexe. Il fonctionne avec un nombre variable de sources. On peut y intégrer une nouvelle source sans qu'il soit nécessaire de disposer d'une longue période d'observation rétrospective. Les données aberrantes sont décelées automatiquement et corrigées, de façon à ne pas perturber les estimations. Les expérimentations, encore peu nombreuses, qui ont été réalisées conduisent à penser que ce système est efficace. Après une phase de mise au point et de rodage, il devrait pouvoir être utilisé en production sans trop de risques, en attendant les résultats du prochain recensement de la population, prévu pour 1999.

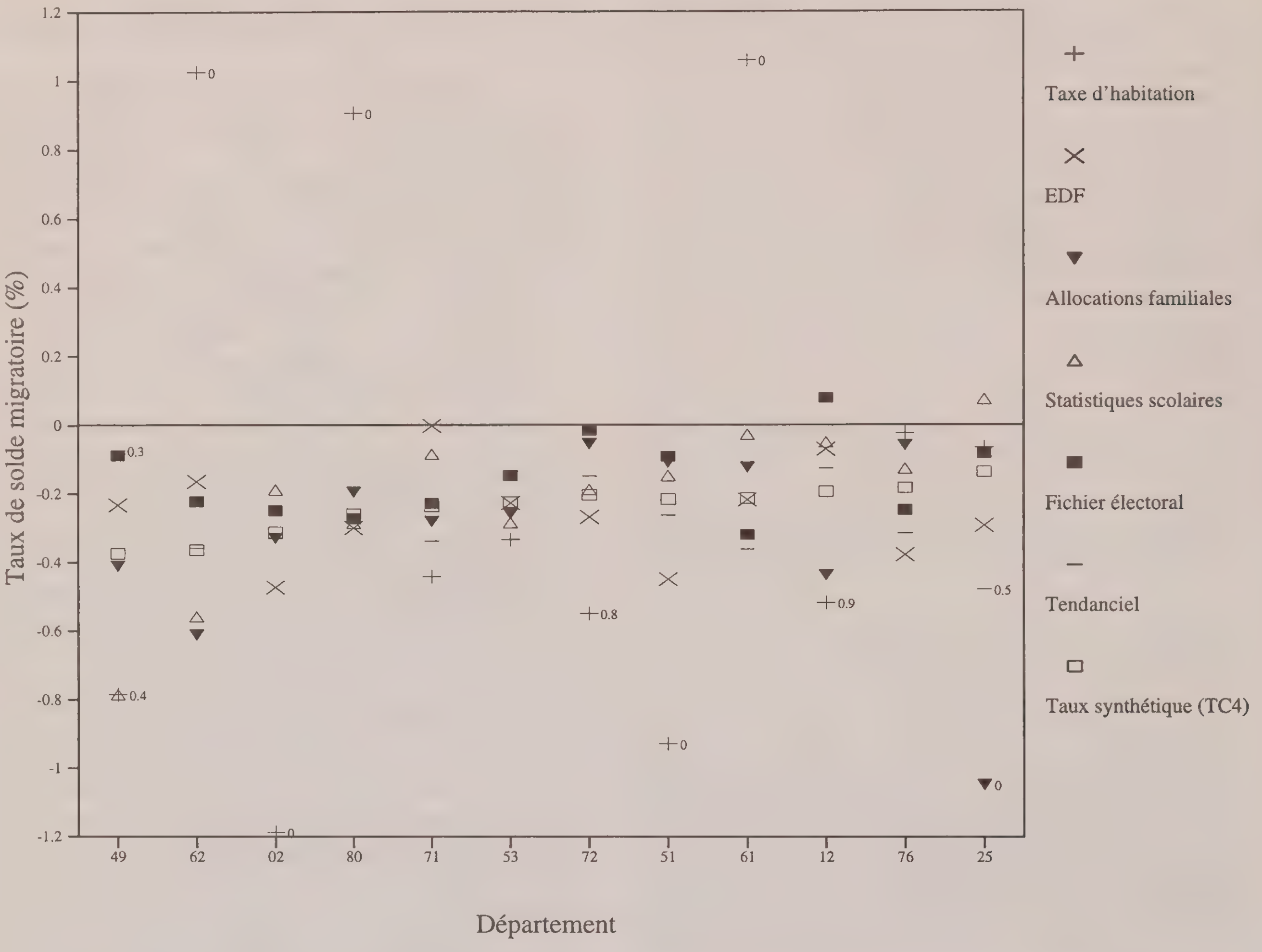


Figure 1. Synthèse des taux de solde migratoire de l'année 1990 pour douze départements, repérés par leur numéro (49, 62...).
N.B.: TC4 est le taux synthétique obtenu après trois itérations. Lorsque le poids d'une source est annulé ou réduit, la valeur du coefficient de modulation (WM3) est indiquée.

Les résultats conduisent à penser que le système est encore plus efficace que ce qu'a indiqué le test rétrospectif sommaire réalisé sur la période intercensitaire 1982-1990 avec les mêmes sources. En effet, en dehors de la source TH, encore perturbée, les estimations provenant des différentes sources sont plus convergentes qu'elles ne l'étaient en moyenne dans le test rétrospectif (*cf.* tableau 4).

Cela n'a d'ailleurs rien d'étonnant, compte tenu du caractère rudimentaire du système testé sur la période intercensitaire 1982-1990. En effet les données utilisées étaient sommaires, voire fragmentaires, en raison de la difficulté à mobiliser en 1993 des données de gestion pour des années anciennes (1982, ...); en outre, les relations utilisées pour tirer de chaque source une estimation du taux de solde migratoire étaient simplistes; enfin, la méthode de synthèse était moins élaborée.

Notons que l'intégration d'autres sources, des données de l'impôt sur le revenu notamment, ne peut que renforcer encore l'efficacité du système.

Tableau 4
Moyenne des écarts dans le test rétrospectif

	TH	EDF	AF	EN	FE
1982	0,26	0,34	0,50	0,47	0,34
1983	0,28	0,33	0,48	0,47	0,32
1984	0,23	0,28	0,40	0,45	0,34
1985	0,24	0,31	0,48	0,44	0,32
1986	0,23	0,33	0,40	0,33	
1987	0,40	0,28	0,41	0,27	
1988	0,84	0,29	0,30	0,37	0,24
1989	0,97	0,21	0,30	0,33	0,35
Moyenne générale	0,43	0,30	0,41	0,39	0,32

Nota:
- Le nombre de taux par année est généralement de 96, sauf pour AF (89) et FE (94).
- La source «fichier électoral» n'a pas fourni de taux pour 1986 ni 1987.
- La source «Taxe d'habitation» a commencé à être perturbée en 1987.
- Les valeurs des écarts correspondent à des taux exprimés en %.

statistique de rang un peu plus élaborée, mais néanmoins simple, compte tenu du petit nombre de valeurs; cette statistique est la moyenne, pondérée respectivement par 1/2, 1/4, 1/4, des trois quartiles:

- la médiane des taux $TC_S(n,z)$ pondérés par les poids a priori W_S,
- le quartile inférieur (Q1) des taux pondérés,
- le quartile supérieur (Q3) des taux pondérés.

(2) Les taux $T1(n,z)$ ainsi obtenus sont calés sur le taux de solde migratoire du niveau supérieur, par simple translation:

$$TC1(n,z) = T1(n,z) + TREF(n) - \sum_z (T1(n,z)P(n,z)) \Big/ \sum_z P(n,z)$$

où $P(n,z)$ est la population de la zone z au 1er janvier de l'an n et $TREF(n)$ le taux de solde migratoire du niveau supérieur (le taux national pour la synthèse départementale).

(3) On calcule, dans chaque zone, les écarts de chaque taux à cette valeur centrale calée:

$$EC1_S(n,z) = |\, TC_S(n,z) - TC1(n,z)\,|.$$

(4) Pour chaque source et chaque zone, l'ampleur de cet écart est appréciée par rapport à la «norme» d'éloignement NO_S propre à la source. Cette «norme» est déterminée empiriquement à partir des données disponibles: c'est en principe la moyenne des écarts constatés dans le passé, anomalies exclues. Il en résulte une première modulation du poids affecté a priori à cette source:

- si $EC1_S(n,z) \le a1\,NO_S$, où $a1$ est un paramètre à choisir (voisin de 2), on ne modifie pas W_S, poids a priori de S. Autrement dit, si $WM1_S(n,z)$ est le coefficient de modulation de W_S (coefficient compris entre 0 et 1), on prend $WM1_S(n,z) = 1$;
- si $EC1_S(n,z) > b1\,NO_S$, où $b1$ est un autre paramètre (voisin de 3), on met W_S à 0, c'est-à-dire qu'on élimine la source S: $WM1_S(n,z) = 0$;
- si $a1\,NO_S < EC1_S(n,z) \le b1\,NO_S$, on interpole $WM1_S(n,z)$ en fonction de la valeur de $EC1_S(n,z)$:

$$WM1_S(n,z) = (b1\,NO_S - EC1_S(n,z))/((b1 - a1)NO_S).$$

(5) A l'issue de cette première phase, on dispose donc de nouveaux poids propres à chaque source et à chaque zone, qui permettent d'éliminer ou de sous-pondérer localement les taux suspects: $W1_S(n,z) = W_S WM1_S(n,z)$.

6.4 Itérations

(1) A l'aide des poids ainsi modifiés $W1_S(n,z)$, on estime pour chaque zone une nouvelle valeur centrale, en prenant cette fois la moyenne pondérée des taux:

$$T2(n,z) = \sum_S (TC_S(n,z)W1_S(n,z)) \Big/ \sum_S W1_S(n,z).$$

(2) On cale chaque taux $T2(n,z)$ sur le taux de solde migratoire du niveau supérieur, par translation. On obtient $TC2(n,z)$.

(3) On calcule, dans chaque zone, les écarts de chaque taux au taux moyen calé: $EC2_S(n,z) = |\, TC_S(n,z) - TC2(n,z)\,|$. À partir de ces écarts, on calcule de nouveaux coefficients de modulation des poids a priori, en utilisant des paramètres $a2$ et $b2$, pouvant être différents de $a1$ et $b1$ (inférieurs en principe). On obtient ainsi de nouveaux poids $W2_S(n,z)$ prenant mieux en compte les anomalies, car celles-ci ont été appréciées par rapport à une meilleure tendance centrale. Avec ces poids, on estime un nouveau taux synthétique $T3(n,z)$, que l'on cale sur le niveau supérieur pour obtenir $TC3(n,z)$.

(4) On répète les opérations du point 3) avec les mêmes paramètres $a2$ et $b2$. Les tests menés au niveau départemental sur 1982-1990 montrent que la convergence est en général rapide; les taux sont très souvent stabilisés à partir de la quatrième itération.

7. MISE EN ŒUVRE AU NIVEAU DÉPARTEMENTAL

Le système d'estimation qui vient d'être présenté dans ses grandes lignes – et qui est destiné à être utilisé de façon opérationnelle pour les années 1990 et suivantes – a été mis en œuvre par la mission pour l'année 1990 au niveau départemental, avec les cinq sources suivantes: taxe d'habitation (TH), abonnés électriques (EDF), allocations familiales (AF), statistiques scolaires (EN), fichier électoral (FÉ), plus l'estimation tendancielle (TEND).

La figure 1 illustre les résultats obtenus pour quelques départements. Le tableau 3 présente les valeurs des poids et des normes retenues pour faire fonctionner le système. Ce tableau présente également certaines statistiques provenant de la synthèse des taux de solde migratoire et portant notamment sur les écarts entre les taux issus de chaque source et les taux synthétiques.

Tableau 3
Mise en œuvre pour l'année 1990 au niveau départemental
Paramètres et statistiques

	TH	EDF	AF	EN	FÉ	TEND
Poids	115	100	80	70	80	100
Norme	0,15	0,17	0,19	0,20	0,19	0,12
Nombre de taux	96	96	89	96	94	96
Moyenne des écarts	0,55	0,14	0,30	0,19	0,14	0,13
Nombre de taux «aberrants»	37	2	17	3	1	6
Moyenne des écarts sans les taux «aberrants»	0,15	0,13	0,16	0,16	0,13	0,11

Nota:
- Coefficients (a; b) appliqués aux normes: (2,5; 3,5) à la première itération, puis (2; 3).
- Les valeurs des écarts et des normes correspondent à des taux exprimés en %.
- Les écarts sont calculés par rapport aux taux synthétiques après trois itérations.
- Les taux «aberrants» sont ceux dont le poids est annulé après trois itérations.

générations l'année d'après (c'est-à-dire de l'effectif des «6-10 ans» l'année $n + 1$) et en défalquant les décès.

Enfants bénéficiaires d'allocations familiales

L'effectif des «0-17 ans» est estimé en supposant qu'il évolue comme le nombre d'enfants bénéficiaires d'allocations familiales. On en déduit un solde migratoire de «jeunes» en comparant cette estimation à l'effectif résultant d'une évolution sans migrations, c'est-à-dire sous le seul effet du mouvement naturel.

6. SYNTHÈSE

6.1 Principes

Les différentes estimations élémentaires du taux de solde migratoire annuel font l'objet d'un traitement statistique, afin d'en tirer un «taux synthétique», retenu comme estimation finale. Le traitement permet d'éliminer les valeurs aberrantes, de sous-pondérer les valeurs suspectes et, plus généralement, d'attribuer à chaque source un poids adapté à ses performances.

Plus précisément, chaque source pouvant «dériver», les différentes estimations élémentaires sont en général biaisées; on les corrige d'abord du biais national de la source correspondante pour l'année considérée, biais qu'on estime au préalable. En procédant ainsi, on suppose implicitement que l'écart entre le biais local et le biais national est de faible importance par rapport au flou irréductible. Lorsqu'on disposera d'estimations pour plusieurs années, on devrait pouvoir tester cette hypothèse, et, le cas échéant, la remplacer par une hypothèse mieux adaptée à la réalité, afin d'améliorer la correction des biais au niveau local.

Notons qu'une opération en apparence aussi simple que la correction du biais national nécessite néanmoins quelques précautions. La solution consistant à opérer un calage brutal sur le taux de solde migratoire national, considéré par définition comme la bonne référence, est peu satisfaisante, en raison des anomalies qui peuvent venir perturber le calage. Il est donc préférable d'estimer les biais au cours d'un processus où l'on élimine aussi les anomalies. Le processus est analogue à celui qui est utilisé pour la synthèse et qui est décrit ci-après. Cependant, la détermination des biais, supposés nationaux et donc calculés sur 96 départements, est moins sensible aux anomalies que celle des taux synthétiques, calculés sur un petit nombre de sources. Seules les anomalies importantes sont susceptibles de fausser sensiblement le calage des taux et doivent donc être corrigées.

Le taux de solde migratoire «synthétique» est une moyenne pondérée des estimations élémentaires ainsi «calées». On attribue à chaque source S un poids «a priori» W_S censé refléter sa précision à moyen terme. Mais de plus, pour une année et une zone données, ce poids est modulé pour prendre en compte le caractère plus ou moins vraisemblable du taux correspondant. Ainsi, un taux «anormalement éloigné» des taux issus des autres sources – en pratique d'une valeur centrale de l'ensemble des taux de la zone – voit son poids annulé ou réduit. Pour cela, on examine l'écart entre le taux provenant de chaque source et la valeur centrale retenue et on le compare à une «norme» d'écart NO_S propre à la source, déterminée empiriquement à partir des données disponibles: si l'écart est inférieur à «a fois» la norme, on ne modifie pas le poids a priori; s'il est supérieur à «b fois» la norme, on met le poids à 0; entre les deux, on multiplie le poids par un coefficient, compris entre 0 et 1, calculé par interpolation.

Notons que l'estimation tendancielle est formellement traitée comme celles provenant des sources exogènes; son poids est annulé lorsqu'elle est considérée comme non vraisemblable, parce que trop éloignée des autres estimations.

La synthèse est réalisée de manière automatique, ce qui assure une homogénéité et une logique explicite aux traitements mis en œuvre. Cela ne supprime pas, pour autant, la nécessité de contrôler les résultats obtenus.

6.2 Présentation théorique

Sur le plan théorique, on a cherché à utiliser les raisonnements et les techniques de l'estimation robuste, exposées par exemple dans Hoaglin, Mosteller et Tukey (1983). La méthode retenue s'inscrit dans le cadre des *M*-estimateurs de tendance centrale et plus précisément dans la catégorie des *W*-estimateurs, qui mettent en œuvre l'algorithme des moindres carrés repondérés.

Les taux de solde migratoire pour l'année n et la zone z issus des différentes sources S (et corrigés de leurs biais nationaux) étant notés $TC_S(n, z)$, le taux synthétique $T(n, z)$ est solution de l'équation implicite:

$$\sum_S W_S \cdot NO_S \cdot \Psi\left(\frac{TC_s(n,z) - T(n,z)}{NO_S}\right) = 0,$$

où la fonction Ψ est de type redescendant à point de rejet fini:

$$\begin{aligned} \Psi(r) &= r && \text{pour } |r| \le a, \\ \Psi(r) &= r\frac{b - |r|}{b - a} && \text{pour } a < |r| \le b, \\ \Psi(r) &= 0 && \text{sinon.} \end{aligned}$$

Un processus itératif permet d'affiner progressivement le traitement automatique des données suspectes.

6.3 Première analyse des distances de chaque taux à la valeur centrale des taux

(1) Pour chaque zone z, on calcule une première valeur centrale des taux «calés» $TC_S(n, z)$. La valeur centrale retenue doit être peu sensible à l'existence éventuelle de valeurs très éloignées pour certaines sources, mais aussi être d'autant plus influencée par une source que cette source est en moyenne plus précise. Dans ces conditions, plutôt que de choisir la médiane – qui répondrait à la première condition – on retient une

Pour chacune des sources expérimentées et jugées «bonnes», au moins au niveau départemental, une méthode est proposée. Les cinq sources retenues sont les suivantes: taxe d'habitation; abonnés électriques; enfants bénéficiaires d'allocations familiales; statistiques scolaires; fichier électoral.

Les données relatives à la composition des foyers fiscaux, figurant dans les fichiers de l'impôt sur le revenu, constituent une sixième source qui devrait fournir de très bons résultats. Cependant, jusqu'à présent, ces données n'ont été analysées que pour quelques départements et la méthode d'utilisation n'est pas encore complètement définie.

Il est proposé en outre d'intégrer au système une estimation tendancielle du taux de solde migratoire.

Deux catégories de méthodes sont utilisées. La première concerne les sources relatives aux ménages; la deuxième celles portant sur des individus.

5.1 Sources relatives aux ménages

Certaines sources fournissent une information sur l'évolution du nombre de ménages. C'est le cas des sources «*taxe d'habitation*» (TH) et «*abonnés électriques*». La taxe d'habitation est un des quatre principaux impôts directs locaux. Comme son nom l'indique, elle s'applique aux logements occupés, selon des modalités différentes pour les résidences principales et les résidences secondaires. C'est la situation au 1er janvier de l'année d'imposition qui est prise en compte. Depuis les années 1980, la source TH est à la base des estimations départementales de population réalisées par l'INSEE (Descours 1992); la source «abonnés électriques» lui a été substituée au début des années 1990, en raison des perturbations provoquées par une modification du système de gestion qui s'est généralisée progressivement à tous les départements.

La méthode retenue pour utiliser ces sources est classique dans son principe. Elle conduit directement à une estimation de la population totale et comporte trois étapes principales:

(1) estimation du nombre de ménages;
(2) estimation de la taille moyenne des ménages et passage à l'estimation de la population des ménages;
(3) ajout de la population «hors ménages».

Dans la première étape, on suppose que le nombre de ménages évolue comme les données fournies par la source (nombre de résidences principales TH ou nombre d'abonnés électriques). La seconde étape est la plus délicate. Elle repose à la fois sur l'utilisation des statistiques de personnes à charge contenues dans les fichiers TH et sur une estimation, de nature tendancielle, de la taille moyenne des ménages.

Dans le système «multi-sources» proposé, on passe au taux de solde migratoire, pour confrontation avec les autres sources, à l'aide des statistiques de l'état civil (*cf.* section 4).

5.2 Sources relatives à des individus

Les autres sources utilisées portent sur des individus. Seule une certaine tranche d'âge X de la population est en général couverte convenablement. La méthode comporte alors deux étapes principales:

(1) estimation, à partir de la source, du taux de solde migratoire de la population d'âge X;
(2) passage au taux de solde migratoire de l'ensemble de la population.

La deuxième étape repose sur la relation statistique suivante, observée dans le passé, entre la variation, d'une période à l'autre, du taux de solde migratoire global (T) et celle du taux de solde migratoire pour la population d'âge X (TX):

$$T_2 - T_1 = \delta_X (TX_2 - TX_1),$$

où δ_X est un coefficient voisin de 1, dépendant de la tranche d'âge X. Cette relation est voisine de celle utilisée par de Guibert-Lantoine (1987) pour estimer la population à partir des statistiques scolaires.

Pour les tranches d'âge correspondant aux différentes sources utilisées, les valeurs, estimées par régression linéaire, du coefficient δ_X(+/−2 écarts-types) sont présentées dans les tableaux 1 et 2.

Tableau 1
Estimation de δ_X sur les départements, hors Corse, soldes internes

Période 1	Période 2	Âge en fin de période		
		0-19 ans	10-14 ans	35 ans ou plus
1962-1968	1968-1975	0,76 (+/- 0,04)	0,69 (+/-0,06)	1,24 (+/-0,09)
1968-1975	1975-1982	0,77 (+/-0,03)	0,88 (+/-0,06)	1,56 (+/-0,08)
1975-1982	1982-1990	0,70 (+/-0,11)	0,49 (+/-0,10)	1,26 (+/-0,17)

Tableau 2
Estimations de δ_X sur le couple de périodes 1975-1982 et 1982-1990, hors Corse, soldes totaux

	Âge en fin de période		
	0-18 ans	9-15 ans	35 ans ou plus
Départements	0,65 (+/-0,11)	0,57 (+/-0,10)	1,22 (+/-0,16)
Département – zone d'emploi	0,65 (+/-0,04)	0,59 (+/-0,04)	1,17 (+/-0,06)

Quant à la première étape, elle dépend de la source:

Fichier électoral

Les migrations électorales annuelles pour la tranche d'âge retenue (les «30 ans ou plus») sont fournies directement par le fichier électoral géré par l'INSEE. On passe du taux de solde migratoire électoral au taux de solde migratoire résidentiel en divisant le premier par un coefficient reflétant l'ampleur de la révision électorale.

Statistiques scolaires

Le solde migratoire des «5-9 ans» est obtenu en soustrayant leur effectif l'année n de celui des mêmes

3. UTILISATION SIMULTANÉE DE PLUSIEURS SOURCES

Pour utiliser conjointement plusieurs sources, différentes méthodes sont envisageables.

Une méthode universelle – et simple à mettre en œuvre – est la *régression multiple*. Sous forme simplifiée, cela revient à utiliser, pour toute zone z, la relation suivante:

$$P(n+1,z)/P(n,z) = c + \sum_S (k_S N_S(n+1,z)/N_S(n,z)),$$

où $P(n,z)$ est la population de la zone z au 1er janvier de l'an n, les $N_S(n,z)$ sont les effectifs provenant de chaque source S à la même date et les k_S des coefficients, qu'on estime par régression multiple sur une période passée. c est ici un terme constant qui ne sert qu'à la régression, le *calage* sur la population nationale permettant de corriger la dérive éventuelle.

Cette méthode est utilisée dans certains pays, le Canada et les États-Unis notamment (voir par exemple Statistique Canada 1987 et Long 1993). Néanmoins, elle n'a pas été retenue car elle présente de nombreux inconvénients:

- il faut pouvoir estimer les coefficients; c'est-à-dire disposer des données de chaque source sur une période passée assez longue;
- les coefficients peuvent évoluer avec le temps, sans qu'on puisse maîtriser cette évolution;
- comme on l'a déjà dit, les sources administratives sont, pour des raisons diverses (changements de réglementation, à-coups de gestion, erreurs...), sujettes à ce qu'on peut appeler des «anomalies». Pour chaque source S, l'importance de ces «anomalies» se reflète en partie dans le coefficient k_S, plus ou moins selon que leur effet à moyen terme a été plus ou moins grand sur la période d'étalonnage; mais les anomalies interviennent néanmoins dans les estimations avec le même poids que les «bonnes» données de la même source. Les estimations sont alors fortement perturbées.

Une autre méthode est celle dite «*composite*». Chaque source sert à estimer la population d'une ou plusieurs classes d'âge: la classe d'âge X bien couverte par la source, mais aussi parfois une autre classe présentant à coup sûr une évolution très voisine de celle de la classe X (par exemple les «30-45 ans», si X représente les «moins de 18 ans»). Il faut alors disposer d'indicateurs appropriés pour les autres composantes de la population et gérer correctement la consolidation de ces estimations «par parties».

Ce genre de méthode, utilisé aux États-Unis (Long 1993), nous a paru problématique, notamment à cause de la difficulté à traiter convenablement les «anomalies».

Le *système «multi-sources» proposé* repose sur une synthèse robuste d'estimations provenant des différentes sources. Il combine un raisonnement démographique et des techniques purement statistiques. Il s'inspire des expériences menées à la Direction régionale de Bretagne de l'INSEE, au début des années 1970 (Laurent et Guéguen 1971, Guéguen 1972). La défaillance de l'une des sources n'empêche pas un tel système de fonctionner, même si ses performances sont un peu dégradées.

4. UNE BASE DÉMOGRAPHIQUE

Le raisonnement démographique qui est à la base du système est élémentaire: en supposant connue la population totale $P(n)$ d'une zone au 1er janvier de l'an n, la population $P(n+1)$ de la zone au 1er janvier de l'an $n+1$ s'en déduit par ajout des deux composantes de la variation au cours de l'année n: l'excédent naturel (naissances moins décès) d'une part, et le solde migratoire (immigrants moins émigrants) d'autre part.

$$P(n+1) = P(n) + N(n) - D(n) + I(n) - E(n).$$

En France, l'excédent naturel est fourni annuellement au niveau communal par les statistiques de l'état civil. Si ces dernières ne sont pas encore disponibles sous forme définitive, ce qui est souvent le cas au troisième trimestre de l'année $n+1$, il est facile de les estimer avec une faible marge d'incertitude.

La seule inconnue est donc le solde migratoire sur l'année n: $SM(n) = I(n) - E(n)$ ou, ce qui est équivalent, le taux de solde migratoire $T(n) = SM(n)/P(n)$. En d'autres termes, estimer la population revient à estimer le solde migratoire depuis la dernière date où cette population est connue (ou supposée telle), et réciproquement.

En France, les soldes migratoires ont une importance non négligeable mais néanmoins modeste par rapport à d'autres pays, comme le Canada ou les États-Unis par exemple. En outre, ils présentent en général une certaine inertie, du moins à des niveaux géographiques relativement agrégés. Une façon d'apprécier l'influence de leurs variations, d'une période intercensitaire à la suivante, consiste à mesurer les erreurs qu'on aurait commises sur chaque période, si on avait estimé les populations en reconduisant les taux de solde migratoire annuels moyens de la période précédente. Sur la période 1982-1990, pour les départements (sans la Corse), l'erreur moyenne en fin de période (en 1990, au bout de huit ans) n'aurait été que de 1,3 %. Il n'était pas sûr, au démarrage de la mission, qu'on puisse atteindre une précision nettement meilleure. Toutefois, en 1975 comme en 1982, l'erreur moyenne qu'on aurait commise, avec la méthode tendancielle, aurait été beaucoup plus forte: 2,8 % et 2,7 % respectivement (sur sept ans). On peut donc penser que la période 1982-1990 a été exceptionnelle et qu'à l'avenir les inflexions redeviendront plus marquées.

5. DES ESTIMATIONS ISSUES DES DIFFÉRENTES SOURCES

On tire de chaque source, par une méthode appropriée, une estimation du taux de solde migratoire annuel de l'ensemble de la population. Les méthodes qui peuvent être utilisées dépendent des données disponibles.

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 99-106
Statistique Canada

Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population en France

GEORGES DECAUDIN et JEAN-CLAUDE LABAT[1]

RÉSUMÉ

La France ne disposant pas de registres de population, les recensements de la population y constituent la base du système d'informations socio-démographiques. Cependant, entre deux recensements, l'actualisation de certaines données est nécessaire, notamment à un niveau géographique fin, d'autant plus que les recensements ont, pour diverses raisons, tendance à s'espacer. Une mission, dont l'objectif était de proposer un système améliorant substantiellement le dispositif d'estimations locales de population en vigueur, a été créée en 1993 au sein de l'Institut National de la Statistique et des Études Économiques. Elle s'est consacrée à une double tâche: réaliser une synthèse efficace et robuste des informations apportées par différentes sources administratives et mobiliser un nombre suffisant de «bonnes» sources. Le système «multi-sources» qu'elle a conçu et qui est présenté ici est souple et fiable, sans être trop complexe.

MOTS CLÉS: Estimations de population; fichiers administratifs; estimation robuste.

1. INTRODUCTION

En France, comme dans tous les pays ne disposant pas de registres de population, les recensements de la population sont la base du système d'informations socio-démographiques. Cependant, ce sont des opérations très lourdes qui, à l'heure actuelle, ne peuvent être réalisées plus fréquemment que tous les sept ou huit ans. Dans l'intervalle, l'actualisation de certaines données est donc nécessaire, notamment à un niveau géographique fin, d'autant plus que les recensements ont, pour diverses raisons, tendance à s'espacer. Ainsi les estimations locales de population constituent un enjeu important pour l'Institut National de la Statistique et des Études Économiques (INSEE).

Malgré les progrès accomplis dans ce domaine, la situation, en 1993, pouvait paraître encore assez peu satisfaisante. Par rapport au recensement de la population de 1990, les estimations de population réalisées, sur la base du recensement précédent (1982), pour les départements métropolitains avaient présenté des écarts parfois importants.

L'INSEE a donc créé une mission à caractère méthodologique, chargée de proposer un système améliorant substantiellement le dispositif en vigueur. Initialement, le prochain recensement devait avoir lieu en 1997. Il semblait donc raisonnable de faire fonctionner le nouveau système de façon expérimentale jusqu'au recensement, afin de vérifier ses performances, avant de l'utiliser en production. Le report du recensement à 1999 a renforcé la nécessité d'aboutir vite, afin de pouvoir utiliser le nouveau système dès 1996.

Pour atteindre son objectif, la mission s'est consacrée, avec le maximum de pragmatisme, à une double tâche: réaliser une synthèse efficace et robuste des informations apportées par différentes sources administratives et mobiliser un nombre suffisant de «bonnes» sources. Le système «multi-sources» qu'elle a conçu, et qui est présenté ici, n'est pas trop complexe et semble efficace. On en trouvera une présentation plus détaillée dans Decaudin et Labat (1996).

2. PRINCIPALES CONCLUSIONS

Les principales conclusions de la mission sont les suivantes:

(1) Il est impossible d'améliorer les estimations de population totale au moyen d'enquêtes par sondage, à moins d'imaginer une enquête d'une taille telle qu'elle s'apparenterait à un recensement.

(2) Aucune source de données administratives ne reflète suffisamment bien les évolutions de population. Toutes les sources peuvent présenter localement des dérives, des ruptures, des à-coups..., qui ne sont pas toujours faciles à déceler. En outre, il est souvent très difficile, voire impossible, d'obtenir de l'organisme responsable, même à l'échelon local, des éléments d'explication et surtout, lorsqu'il s'agit d'une erreur, les éléments de correction. De toute façon, il est imprudent de se fonder sur une seule source administrative, aussi bonne soit-elle, car sa pérennité n'est jamais assurée.

(3) En revanche, il est possible d'améliorer substantiellement les estimations de population totale en utilisant simultanément plusieurs sources. Un système «multi-sources», analogue à celui présenté ici mais plus rudimentaire, a été testé rétrospectivement, sur la période intercensitaire 1982-1990, pour les 96 départements métropolitains. L'erreur moyenne (moyenne des écarts relatifs en valeur absolue avec les résultats du recensement de mars 1990) est descendue au-dessous de 0,9 %, alors que l'erreur moyenne commise à l'époque, avec le système d'estimation en vigueur, était de 1,4 %.

[1] Georges Decaudin et Jean-Claude Labat, Institut National de la Statistique et des Études Économiques, 18, Boulevard Adolphe-Pinard, 75675 Paris, CEDEX 14.

que $E_2[(\sum_{i\in F_h^*} w_i\, z_i)^2]/n_h \approx (\sum_{i\in F_h^*} w_i y_i)^2/n_h$. L'équation (14) provient de cette quasi-égalité et sur les équations (11) et (12) n_h étant élevé, $n_h/(n_h - 1) \approx 1$).

Contre-exemples de l'estimateur jackknife de l'estimateur de développement double

Comme contre-exemple de la forme répétée de l'équation (16), prenons le cas où chaque grappe ne renferme qu'un élément, $H = G = 1$, et où y_i est toujours égal à 1. Dans ce cas, $t_3 = T$, et t_3 n'a pas de variance. Malheureusement, $t_{(1j)3} = T[n_1/(n_1 - 1)](m - 1)/m$ quand $j \in s$ et $Tn_1/(n_1 - 1)$ dans les autres cas. Donc, $(t_{(1j)3} - T)/T = O_p(1/m)$. Le rapport v_{J3}/T^2 qui dérive de $t_{(1j)3}$ serait lui aussi égal à $O(1/m)$, puisqu'il s'agit de la somme des termes n_1 d'ordre $O(1/m^2)$.

Bien que v_{J3}/T^2 corresponde à $O(1/m)$, v_{J3} ne se rapproche pas assez de zéro pour nous être utile. En effet, si y_i était toujours égal à $N(1,1)$, la variance relative de t_3 serait $1/m$, qui correspond aussi à $O(1/m)$. Pour que v_{J3} soit presque égal à zéro, v_{J3}/T^2 devrait donc être inférieur à $O(1/m)$. Cela n'étant pas le cas, l'estimateur de variance jackknife est loin de ne pas être biaisé.

Comme contre-exemple de la forme répétée de l'équation (17), examinons le cas où chaque grappe renferme de nouveau un seul élément et où y_i est égal à un, mais où $H = m$, $G = 1$, la population de h est toujours égale à N_0, $n_h = 2$ pour toutes les valeurs de h, et $M_1 = 2m$. Il s'ensuit que $T = t_3 = mN_0$, si bien que t_3 ne présente pas de variance. La répétition $t^*_{(hj)3}$ peut donc prendre quatre valeurs. Si $hj \in s$ et $hj' \in s\,(j \neq j')$, alors, $t^*_{(hj)3} = [(m/2)(2m - 1)/(m - 1)]N_0$. Si $hj \in s$ et $hj' \notin s$, alors, $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/(m - 1)]N_0$. Si $hj \notin s$ et $hj' \in s$, alors, $t^*_{(hj)3} = [(m/2)(2m - 1)/m]N_0$. Si $hj \notin s$ et $hj' \notin s$, alors, $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/m]N_0$. Dans aucun de ces cas, $(t^*_{(hj)3} - T)/T = O_p(1/m)$, de sorte que l'estimateur de variance jackknife ne peut être presque dépourvu de biais.

Estimateur de régression de la deuxième phase

Pour étayer l'argumentation sur l'estimateur de régression de l'équation (21), supposons que le plan d'échantillonnage et la population soient tels qu'on obtient confirmation des relations asymptotiques que voici. En premier lieu,

$$\sum_{i\in S} w_i \boldsymbol{x}_i \Big(\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i\Big)^{-1} d_i \boldsymbol{x}_i' - 1 = O_p(1/\sqrt{m}), \qquad \text{(A6)}$$

qui est une généralisation de l'équation (A1). De même, les équations (A2) et (A3) peuvent être généralisées pour donner

$$\sum_{i\in S_g} w_{hji} d_i q_i \Big/ \sum_{i\in S_g} w_i d_i q_i - 1 = O_p(1/m), \qquad \text{(A7)}$$

et

$$\sum_{i\in s_g} w_{hji} e_i d_i q_i \Big/ \sum_{i\in s_g} w_i e_i d_i q_i - 1 = O_p(1/m) \qquad \text{(A8)}$$

pour toutes les valeurs de q_i, où q_i est un élément de la matrice $\boldsymbol{x}_i' \boldsymbol{x}_i$. Enfin, l'équation (A4) se généralise pour devenir

$$\sum_{i\in S_g} w_{hji} d_i p_i \Big/ \sum_{i\in S_g} w_i d_i p_i - 1 = O_p(1/m) \qquad \text{(A9)}$$

pour toutes les valeurs de p_i, où p_i représente un élément de la matrice $\boldsymbol{x}_i' y_i$.

BIBLIOGRAPHIE

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOVAČEVIĆ, M.S., et YUNG, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique. *Techniques d'enquête*, 23, 47-59.

KREWSKI, D., et RAO, J.N.K. (1981), Inferences from stratified samples: properties of linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. *Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies*, (Éds. W.G. Madow, I. Olkin, et D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4, 811-822.

RAO, J.N.K., et WU, C.F.J. (1985). Inferences from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1, 381-397.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada:* 1984-1990. N° 71-526 au catalogue, Statistique Canada.

STUKEL, D.M., et BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Direction de la méthodologie, document de travail, SSMD, 92-009E, Statistique Canada.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

et que, pour *tout* échantillon de première phase,

$$\left(\sum_{k\in S_g} w_k \Big/ \sum_{k\in s_g} w_k\right)(m_g/M_g) - 1 = O_p(1/\sqrt{m}) \qquad \text{(A1)}$$

pour toutes les valeurs de g. Ces hypothèses justifient l'équation (5) dans le corps du texte.

L'analyse présume que G est borné et que chaque valeur de m_g présente le même degré asymptotique que m. La chose n'est réalisable que lorsqu'on définit S_g *après* prélèvement de l'échantillon de la première phase. Sans cela, M_g équivaudrait à une variable aléatoire, si bien qu'on ne pourrait garantir une valeur minimale pour m_g, pour tous les échantillons envisageables de la première phase. En principe, on suppose qu'un mécanisme permet de déterminer S_g et les fractions de l'échantillonnage de la deuxième phase, compte tenu d'un échantillon quelconque de la première phase. Les valeurs exactes de G et de m_g, en revanche, peuvent être arrêtées avant le prélèvement de l'échantillon de la première phase, sans que cela soit toutefois une obligation.

Remarque au sujet du cadre asymptotique

Nous avons montré que l'estimateur jackknife intègre une composante permettant d'estimer la variance à la deuxième phase $E_2[(t_2 - t_1)^2]$) sans introduire de biais asymptotique, *quel que soit* l'échantillon de la première phase (voir l'équation (14)). Par voie de conséquence, cette composante permet d'estimer la variance moyenne à la deuxième phase (donc non conditionnelle) de tous les échantillons possibles de la première phase $E_1\{E_2[(t_2 - t_1)^2]\}$, sans introduire de biais asymptotique.

Nous nous sommes éloignés du cadre décrit ci-dessus dans le travail empirique afin que les résultats soient plus faciles à résumer. Plus précisément, nous avons défini S_g au préalable et laissé M_g varier. Advenant le cas où l'échantillon de la première phase donnerait une valeur M_g inférieure à la valeur m_g désirée (50, par exemple) à la strate de la deuxième phase, nous avions l'intention de retenir tous les sujets de S_g pour constituer l'échantillon de la deuxième phase. La présence de cette strate g de la deuxième phase n'augmenterait donc pas l'erreur quadratique moyenne (ou biais) de t_2 et les hypothèses asymptotiques sur m_g s'avéreraient superflues. Ainsi qu'on a pu le voir, M_g n'a jamais obtenu une valeur inférieure à 50 dans la simulation. Quoi qu'il en soit, on disposait d'une règle applicable aux fractions de l'échantillonnage de la deuxième phase, pour tous les échantillons de la première phase.

Répétitions de l'estimateur jackknife

Deux cadres asymptotiques distincts (au moins) s'appliquent à l'échantillon de la première phase. Le premier comprend un nombre arbitrairement élevé de strates à la première phase, la taille de chacune étant limitée; bref, pour chacune d'elles, $1/n_h = O(1)$ tandis que $1/H = O(1/m)$. Dans le deuxième cas, toutes les strates de la première phase sont arbitrairement importantes, soit $1/n_h = O(1/m)$. On suppose que chaque grappe renferme $O(1)$ éléments dans les deux cas, bref que chaque grappe est bornée.

Puisque m_g est du même ordre asymptotique que m, il est raisonnable de penser que dans l'un ou l'autre cas, pour un échantillon donné de la première phase,

$$\sum_{i\in S_g} w_{hji} \Big/ \sum_{i\in S_g} w_i - 1 = O_p(1/m), \qquad \text{(A2)}$$

$$\sum_{i\in s_g} w_{hji} \Big/ \sum_{i\in s_g} w_i - 1 = O_p(1/m), \qquad \text{(A3)}$$

ce dont on peut se servir pour dériver l'équation (9). De même, on présume que pour tout échantillon de la première phase

$$\sum_{i\in S_g} w_{hji} y_i \Big/ \sum_{i\in S_g} w_i y_i - 1 = O_p(1/m), \qquad \text{(A4)}$$

ce qui donne $r_{hji} - r_i = O_p(1/m)$.

Équations (12), (13) et (14)

Le nombre d'éléments dans chaque grappe étant limité, par B par exemple, le troisième terme de l'équation (12) compte au plus GB^2 termes, un nombre fini.

Chaque terme est d'ordre $1/m_g$ (plus exactement, la probabilité qu'un terme soit asymptotiquement d'ordre supérieur à $1/m_g$ est égale à zéro). Par conséquent, on peut négliger la deuxième ligne de l'équation (12), sur le plan asymptotique.

L'équation (14) se vérifie chaque fois que $1/n_h = O(1)$, car si n_h est inférieur à C (par exemple), le troisième terme de droite de l'équation (13) correspond à la somme d'un maximum de $G(BC)^2$ termes, un nombre fini. Cette fois encore, chaque terme est d'ordre $1/m_g$. On peut donc ignorer la deuxième ligne de l'équation (13) sur un plan asymptotique.

Supposons d'autre part, que chaque rapport $1/n_h$ soit égal à $O(1/m)$. On présumera que le plan d'échantillonnage et la population sont tels que, pour un échantillon quelconque à la première phase,

$$A_h = \sum_{i\in F_h^*} w_i(e_i c_i - 1) r_i \Big/ \sum_{i\in F_h^*} w_i y_i = O_p(1/\sqrt{m}) \qquad \text{(A5)}$$

pour toutes les valeurs de h. Il s'agit d'une hypothèse raisonnable puisque, conditionnellement à l'échantillon de la première phase, le dénominateur de A_h représente le total d'un domaine – soit la somme de $w_i y_i$ pour les éléments de F_h^*. Par conséquent, il correspond à $O(m)$ (sans perte de généralité, on peut supposer que chaque w_i est égal à $O(1)$). Le numérateur de A_h indique l'écart entre l'estimateur de développement (somme des éléments $w_i e_i c_i r_i$ dans F_h^*) d'un échantillon aléatoire simple stratifié et sa cible (la somme des éléments $w_i r_i$ dans F_h^*). L'équation (A.5) repose sur la modeste hypothèse que le plan d'échantillonnage et la population donnent une différence de $O_p(\sqrt{m})$ pour tous les échantillons envisageables de la première phase.

En vertu de l'hypothèse (A5), $\sum_{i\in F_h^*} w_i z_i = \sum_{F_h^*} w_i y_i (1 + A_h)$ équivaut à peu près à $\sum_{i\in F_h^*} w_i y_i$, si bien

ne s'avère pas très difficile. Supposons cependant que le plan d'échantillonnage compte plus de deux phases ou qu'on désire estimer le quotient de deux totaux. Quoi qu'elle demeure réalisable en pareil cas, la linéarisation gagne de plus en plus en difficulté. Il n'en va pas autant avec l'estimateur jackknife.

On peut aisément généraliser les résultats de la partie 3 pour un échantillonnage à p-phases par induction. La lettre h désigne toujours les strates de la première phase, mais la lettre g correspond désormais à celles de la phase p-ième représente le jeu d'éléments de l'échantillon de la S_g phase de la strate g, alors que s_g est le sous-échantillon de la p-ième phase de g. On remplace la valeur w_i de l'équation (2) par a_i de (3), pour l'estimateur de la (p-1)-ième phase. De même, on calcule la valeur $t_{(hj)2}$ de l'estimateur jackknife avec a_{hji} de la (p-1)-ième phase, au lieu de w_{hji}.

Remplacer l'échantillon en grappes stratifié prélevé à la première phase par un échantillon stratifié à plusieurs phases s'avère aussi assez simple (nous laissons au lecteur le soin de le faire). On obtient encore les résultats de la partie 3 pourvu que l'échantillon à plusieurs phases soit toujours prélevé par tirage non exhaustif à la première phase.

Enfin, il n'est pas difficile d'étendre les résultats de la partie 3 à des estimateurs plus complexes. Soit $\boldsymbol{U}_2$, un vecteur des estimateurs de l'équation (2) adoptant la forme t_2. L'erreur quadratique moyenne d'un estimateur quelconque $\boldsymbol{\Theta} = \boldsymbol{g}(\boldsymbol{U}_2)$, où g est une fonction continue, peut être estimée presque sans biais grâce à l'estimateur jackknife, chaque fois qu'on peut en faire autant pour les éléments de U_2. Cette remarque respecte les preuves données dans les ouvrages. Ainsi, Rao et Wu (1985) examinent le plan asymptotique où toutes les valeurs n_h sont bornées, tandis que Wolter (1985; chapitre 4.5) analyse le cas où n_h augmente considérablement de façon arbitraire.

6.2 Régression à la deuxième phase

On peut généraliser l'estimateur t_2 par l'estimateur de régression:

$$t_{2\text{reg}} = \sum_{i \in S} w_i \boldsymbol{x}_i \left(\sum_{i \in s} w_i e_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i \in s} w_i e_i d_i \boldsymbol{x}_i' y_i \right), \quad (21)$$

où S représente l'échantillon original; $\boldsymbol{x}_i$, un vecteur ligne; d_i, une grandeur scalaire et où il existe un vecteur ligne $\boldsymbol{\gamma}$ tel que $d_i \boldsymbol{\gamma} \boldsymbol{x}_i' = 1$ pour toutes les valeurs de i. Dans la pratique, d_i est habituellement égal à 1 pour toutes les valeurs de i. Une exception survient fréquemment quand $\boldsymbol{x}_i = x_i$ et $d_i = 1/x_i$. Dans l'équation (2), $d_i = 1$ pour toutes les valeurs de i, et $\boldsymbol{x}_i$ correspond à un vecteur G de valeur 1 à la g-ième position mais de valeur nulle ailleurs, pour $i \in S_g$.

Soit

$$r_i = y_i - \boldsymbol{x}_i \left(\sum_{i \in S} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i \in S} w_i d_i \boldsymbol{x}_i' y_i \right).$$

La répétition $t_{2\text{reg}(hj)}$ a une forme identique à $t_{2\text{reg}}$, mais w_{hji} est remplacé par w_i. De même, r_{hji} a la même forme que r_i, si ce n'est que w_{hji} se substitue à w_i. Remarquons que e_i ne change pas dans $t_{2\text{reg}}$ et $t_{2\text{reg}(hj)}$.

Puisqu'il n'y a pas eu modification du plan d'échantillonnage, l'équation (6) ne change pas, si ce n'est que désormais $(\sum_{i \in S_g} w_i r_i)^2$ est non négatif au lieu d'être strictement égal à zéro. L'intéressé pourra s'assurer que les équations (10) à (13) gardent leur forme actuelle. Dans l'équation (14), on note que, si biais de l'estimateur jackknife il y a, celui-ci tend (approximativement) à la hausse. Bref, il s'agit d'un estimateur *conservateur* de la variance. Encore une fois, le lecteur est prié de se reporter à l'annexe (équations (A6) à (A9)) pour se faire une meilleure idée des hypothèses asymptotiques.

Le biais de l'estimateur jackknife disparaît quand $\sum_{i \in S_g} w_i r_i = 0$ pour toutes les valeurs de g. Pareille situation survient lorsqu'il existe G vecteurs de rangée $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_G$ de sorte que $d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' = 1$ quand $i \in S_g$ et 0 prend la valeur nulle dans les autres cas (puisque $\sum_{i \in s} w_i r_i = \sum_{i \in s} d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' w_i r_i = \boldsymbol{\gamma}_g \sum_{i \in s} w_i d_i \boldsymbol{x}_i' r_i = \boldsymbol{\gamma}_g \{\sum_{i \in s} w_i d_i \boldsymbol{x}_i' (y_i - \boldsymbol{x}_i [\sum_{i \in s} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i]^{-1} \sum_{i \in s} w_i d_i \boldsymbol{x}_i' y_i)\} = 0$). L'existence de $\boldsymbol{\gamma}_g$ quand $d_i = 1$, signifie qu'un membre de $\boldsymbol{x}_i$ est une variable indicatrice égale à 1 lorsque $i \in S_g$ et a la valeur nulle dans les autres cas, ou qu'un membre de la transformation linéaire de $\boldsymbol{x}_i$ est cette variable indicatrice.

7. CONCLUSION

Notre article avait principalement pour but de montrer qu'un simple estimateur de variance jackknife peut être presque dépourvu de biais lorsque la méthode d'estimation s'articule sur un échantillonnage à deux phases, pourvu qu'on recoure à un estimateur de développement repondéré plutôt qu'à un estimateur de développement double. L'application pratique des résultats théoriques de l'estimateur de développement repondéré dépendra du contexte puisque ces résultats reposent sur une argumentation asymptotique. L'étude de simulation de Monte Carlo que nous avons effectuée donne néanmoins à penser que l'estimateur jackknife a son utilité pour estimer la variance de l'estimateur de développement repondéré, même en présence de strates étonnamment peu importantes à l'échantillonnage de la deuxième phase, c'est-à-dire de strates qui ne comptent que 5 ou 10 éléments.

ANNEXE

Cohérence de l'estimateur de développement repondéré au niveau du plan d'échantillonnage

Pour vérifier la cohérence théorique de t_2 dans l'équation (2), on suppose simplement que le plan d'échantillonnage et la population de y_i sont tels que

$$\left\{ \sum_{g=1}^{G} (M_g / m_g) \sum_{i \in s_g} w_i y_i / T \right\} - 1 = O_p(1/\sqrt{m}),$$

total de personnes occupées selon l'équation (19), tandis que le tableau 2B en fait autant pour le taux d'emploi. Commençons par examiner le premier. On se rend compte que la variance de l'estimateur intégral de la première phase est presque totalement dépourvue de biais (0,94 %). L'estimateur jackknife donne de bons résultats avec l'estimateur de développement repondéré alors que la variance ne présente qu'un léger biais négatif, toujours inférieur à −6 %. Ce biais a tendance à devenir plus négatif (même si ce n'est pas de façon uniforme) à mesure que la taille de l'échantillon de la deuxième phase diminue.

Les deux variantes de l'estimateur jackknife de l'estimateur de développement double, en revanche, donnent de piètres résultats, avec un fort biais positif pour la variance, allant de 46,35 % à 1997,51 %! La deuxième variante est pire que la première, mais les deux, se comportent d'une manière absolument inacceptable.

Le tableau 2B reprend l'analyse pour l'estimation par quotient du taux d'emploi. Les résultats ont de quoi surprendre. En effet, tous les estimateurs de variance se comportent raisonnablement bien, sauf la variante 2 de l'estimateur de développement double quand $m_g = 5$. Outre ce cas, où il atteint 30,46 %, le biais est inférieur à 10 % en valeur absolue.

Dans l'ensemble, les tableaux 2A et 2B appuient fortement l'usage de l'estimateur de variance jackknife avec l'estimateur de développement repondéré, même avec un très petit échantillon de la deuxième phase. Par contre, cet estimateur échoue lamentablement avec l'estimateur de développement double quand on estime les totaux. Il arrive cependant que la variante 1 donne des résultats acceptables, selon l'estimateur et les données.

Bien que la majorité des études insistent sur le *biais* des estimateurs de variance, il vaut la peine d'examiner le *coefficient de variation* des estimateurs de variance pour établir la stabilité des estimations de la variance. Le coefficient de variation estimé (en pour cent) se rapportant au nombre total de personnes occupées et au taux d'emploi apparaît respectivement aux tableaux 3A et 3B. L'expression sous la racine carrée du numérateur à l'équation (20) donne l'EQM de la variance, composée de la valeur quadratique du biais de la variance et de la variance de la variance. Les tableaux 3A et 3B ne présentent pas les valeurs correspondantes des entrées des tableaux 2A et 2B (signalées par un *) pour lesquelles le biais de la variance est trop élevé (supérieur à 20 %, par exemple), car il est clair que ces valeurs seront elles aussi trop élevées. Au tableau 3A, les coefficients de variation estimés associés à l'estimateur de développement repondéré fluctuent entre 46,86 % et 53,42 %, ce qui est caractéristique aux estimateurs de la variance. Des coefficients de variation aussi importants ont été relevés dans d'autres études de simulation sur la variance, notamment celle de Kovačević et Yung (1997). En l'occurrence, on remarquera que les coefficients de variation estimés des estimateurs intégraux de la première phase se situent dans la même fourchette de valeurs. En réalité, ils dépassent légèrement ceux des estimateurs de la deuxième phase.

Tableau 3A
Coefficient de variation de la variance jackknife du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	51,33	49,3	46,86	53,42
EED (Variante 1)	–	*	*	*	*
EED (Variante 2)	–	*	*	*	*
EIPD	56,71	–	–	–	–

Tableau 3B
Coefficient de variation de la variance jackknife du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	59,28	65,66	74,26	103,06
EED (Variant 1)	–	59,24	66,16	72,89	99,1
EED (Variant 2)	–	60,94	73,2	92,71	*
EIPD	78,42	–	–	–	–

EER - Estimateur de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)
La variante 1 utilise la répétition de l'estimateur jackknife de l'équation (16).
La variante 2 utilise la répétition de l'estimateur jackknife de l'équation (17).

Si on les examine un à un, on se rend compte que les coefficients de variation de la variance du taux d'emploi estimé qui apparaissent au tableau 3B sont plus élevés que les coefficients correspondants du tableau 3A. D'autre part, tous les estimateurs se remarquent par une hausse appréciable du coefficient de variation correspondant quand la taille de l'échantillon de la deuxième phase diminue. L'effet est plus prononcé pour les estimateurs par quotient que pour les estimateurs du total. Les coefficients de variation très importants de la colonne $m_g = 5$ aux deux tableaux ne surprendra personne puisque la taille globale de l'échantillon de la deuxième phase (25) est en fait inférieure au nombre d'UPE prélevées à la première phase de l'échantillonnage (36). Le nombre de membres de la population active échantillonnés (c.-à-d. au dénominateur) constitue d'ailleurs un meilleur dénombrement de l'échantillon pour l'estimateur par quotient. Cette valeur varie d'un échantillon à l'autre et est souvent considérablement inférieure à 25.

6. EXTENSION DE L'ESTIMATEUR DE DÉVELOPPEMENT REPONDÉRÉ

6.1 L'estimateur de développement repondéré

Élaborer un estimateur de variance linéarisé pour l'estimateur de développement repondéré de l'équation (2)

Le biais relatif en pour cent de l'estimateur de variance jackknife par rapport à l'erreur quadratique moyenne réelle est estimé par

$$\mathrm{BRP}[v_{Jf}(t^*)] = (\{E_M[v_{Jf}(t^*)] - \mathrm{EQM}_{\text{vraie}}\}/\mathrm{EQM}_{\text{vraie}}) \times 100, \qquad (19)$$

où

$$E_M[v_{Jf}(t^*)] = (1/4\ 000) \sum_{r=1}^{4\ 000} v_{Jfr}(t^*),$$

$$\mathrm{EQM}_{\text{vraie}} = (1/4\ 000) \sum_{r=1}^{4\ 000} (t_r^* - T_y)^2,$$

et $v_{Jfr}(t^*)$ est la valeur de $v_{Jf}(t^*)$ de l'échantillon r.

Le coefficient de variation (en pour cent) de l'estimateur de variance jackknife par rapport à l'EQM/réelle est estimé par:

$$\mathrm{CV}[v_{Jf}(t^*)] = (\{(1/4\ 000)\sum [v_{Jfr}(t^*) - \mathrm{EQM}_{\text{vraie}}]^2\}^{½}/\mathrm{EQM}_{\text{vraie}})\times 100; \qquad (20)$$

bref, la racine de l'erreur quadratique moyenne estimée de l'estimateur de variance, divisée par l'EQM réelle estimée et exprimée en pourcentage.

5.2 Résultats de l'étude

Le tableau 1A indique le biais relatif estimé en pour cent des trois estimations ponctuelles du nombre total de personnes occupées selon l'équation (18). Le tableau 1B en fait autant mais pour le taux d'emploi. Tous les biais ont une valeur absolue inférieure à 1 %.

Tableau 1A
Biais relatif en pour cent des estimations ponctuelles du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	0,14	−0,3	−0,29	−0,56
EED	–	0,16	−0,01	0,03	0,115
EIPD	0,04	–	–	–	–

Tableau 1B
Biais relatif en pour cent des estimations ponctuelles du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	−0,09	−0,31	−0,19	−0,26
EED	–	−0,08	−0,27	−0,12	−0,13
EIPD	−0,09	–	–	–	–

EER - Estimation de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)

Ni l'estimation de Monte Carlo de l'erreur quadratique moyenne (c.-à-d. la valeur $\mathrm{EQM}_{\text{vraie}}$ ni les coefficients de variation correspondants, obtenus grâce à l'estimateur de développement double ou à sa version repondérée, n'apparaissent dans les tableaux car l'article porte principalement sur *l'estimation* de l'erreur quadratique moyenne. L'erreur quadratique moyenne (et les coefficients de variation) qui dérive de l'application des deux estimateurs est comparable, peu importe la taille de l'échantillon (l'écart relatif entre les coefficients de variation correspond à peu près à la moitié de l'écart relatif entre les erreurs quadratiques moyennes). L'estimateur de développement repondéré s'avère légèrement plus efficace lorsqu'il s'agit d'estimer le nombre total de personnes occupées (à savoir, quand $m_g = 5$, l'erreur quadratique moyenne de l'estimateur de développement double augmente de 17 %). Quand on estime le taux d'emploi, l'écart entre l'erreur quadratique moyenne des deux méthodes est inférieur à 1 %. On ne sera guère surpris d'apprendre que l'erreur quadratique moyenne des estimateurs augmente à mesure que la taille de l'échantillon de la deuxième phase diminue.

Tableau 2A
Biais relatif en pour cent de l'estimateur de variance jackknife du nombre total de personnes occupées

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	−0,99	−2,51	−5,81	−5,13
EED (Variante 1)	–	46,35	68,24	78,18	86,22
EED (Variante 2)	–	101,59	278,44	654,99	1997,51
EIPD	0,94	–	–	–	–

Tableau 2B
Biais relatif en pour cent de l'estimateur de variance jackknife du taux d'emploi

Estimateur	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
EER	–	−3,53	−3,45	−7,09	−6,55
EED (Variante 1)	–	−2,46	−1,53	−5,21	−7,41
EED (Variante 2)	–	−0,36	4,91	9,09	30,46
EIPD	2,08	–	–	–	–

EER - Estimateur de développement repondéré (t_2)
EED - Estimateur de développement double (t_3)
EIPD - Estimateur intégral du premier degré (t_1)
La variante 1 utilise la répétition de l'estimateur jackknife de l'équation (16).
La variante 2 utilise la répétition de l'estimateur jackknife de l'équation (17).

Le tableau 2A présente le biais relatif estimé en pour cent de l'estimateur de variance jackknife pour le nombre

5. ÉTUDE DE SIMULATION DE MONTE CARLO

5.1 Conception de l'étude

Les résultats qu'on a pu examiner jusqu'ici sont asymptotiques. Nous avons effectué une étude de simulation de Monte Carlo afin d'évaluer la précision de l'estimateur jackknife en tant qu'estimateur de la variance de l'estimateur de développement repondéré dans un univers fini. Parallèlement, nous avons évalué la précision des deux estimateurs jackknife proposés pour l'estimateur de développement double à la partie 4.

Nous nous sommes servis des données de l'Enquête sur la population active (EPA) canadienne de décembre 1990 pour la province de Terre-Neuve. De cette population finie, nous avons tiré des échantillons répétitifs. L'EPA est la plus vaste enquête-ménage par sondage poursuivie en permanence par Statistique Canada. Les données sur le marché du travail sont recueillies mensuellement grâce à un plan d'échantillonnage complexe à degrés multiples, comportant plusieurs niveaux de stratification. On trouvera plus de précisions sur ce plan d'échantillonnage avant la modification qu'il a subie en 1991 dans Singh, Drew, Gambino et Mayda (1990), ainsi que dans Stukel et Boyer (1992). En bref, les provinces sont stratifiées en «régions économiques», vastes régions à structure économique analogue; Terre-Neuve en compte quatre. Les régions économiques sont subdivisées en strates de niveau inférieur. À Terre-Neuve, le niveau de stratification le plus bas donnait 45 strates comprenant chacune moins de six grappes ou *unités primaires d'échantillonnage* (UPE), ce qui était insuffisant pour l'échantillonnage dans le cadre de la simulation. On a donc regroupé les 45 strates en 18, comprenant chacune 6 à 18 UPE. Les régions économiques ont été préservées lors du regroupement des strates, tout comme on a maintenu les régions métropolitaines de recensement de St. John's et de Cornerbrook.

Dans le cadre de l'étude de Monte Carlo, on a prélevé $R = 4\ 000$ échantillons de la «population» de Terre-Neuve (composée de 9 152 individus), selon le plan d'échantillonnage à deux phases que voici: on a d'abord tiré deux UPE de chaque strate de la première phase par échantillonnage aléatoire simple (EAS) *non exhaustif* (NE). On a ainsi obtenu au total 36 UPE. Tous les ménages des UPE sélectionnées à la première phase (et les personnes composant ces ménages) ont été retenus, ce qui a donné un échantillon en grappes exhaustif à la première phase. À la deuxième phase, tous les éléments précédemment sélectionnés (les sujets, en comptant chaque personne choisie deux fois dans une UPE comme deux sujets distincts) ont été restratifiés en cinq groupes d'âge (< = 14, 15-24, 25-44, 45-64, > = 65) et les éléments de l'échantillon de la deuxième phase (lire les sujets) ont été prélevées par EAS *exhaustif* dans chacune des cinq strates de la deuxième phase.

Nous avons varié la taille de l'échantillon des strates à la deuxième phase en prenant m_g = 5, 10, 20, et 50, de manière à obtenir des échantillons au deuxième degré de taille m = 25, 50, 100 et 250. Quand le nombre de sujets échantillonnés à la première phase faisant partie d'une strate à la deuxième phase était inférieur à la valeur m_g désirée, notre intention était d'établir $m_g = M_g$, mais le cas ne s'est jamais présenté.

Une populaire règle heuristique applicable à «l'estimateur par le quotient distinct» comme l'estimateur de développement repondéré de l'équation (2) est que chaque strate de la deuxième phase comporte au moins 20 éléments (lire, à ce sujet, Särndal, Swensson et Wretman 1992, p. 270). Notre but, en attribuant les valeurs 5 et 10 à m_g, était de vérifier l'utilité d'une telle règle.

Nous avons envisagé deux paramètres intéressants: T_y, soit le nombre total de personnes occupées, et T_y / T_z le taux d'emploi. Dans le cas présent, $T_y = \sum_{i \in U} y_i$, ou $y_i = 1$ quand le sujet i a un emploi et a la valeur nulle dans les autres cas. De même, $T_z = \sum_{i \in U} z_i$, ou $z_i = 1$ quand le sujet i fait partie de la population active (c.-à-d. travaille ou chôme) et a la valeur nulle dans les autres cas. Pour chacun des $R = 4\ 000$ échantillons, nous avons calculé l'estimateur de développement repondéré (EER) t_2 de l'équation (2), l'estimateur de développement double (EED) t_3 de l'équation (15) et l'estimateur de développement intégral de la première phase (EEIPD) t_1 de l'équation (1). Quoique ces estimateurs soient définis en fonction d'un total (nombre de personnes occupées), il est facile d'en étendre l'application à un rapport de totaux (au taux d'emploi, par exemple).

Pour chacun des $R = 4\ 000$ échantillons de la deuxième phase, nous avons calculé la variance jackknife qui correspondait à l'estimateur de développement repondéré et à l'estimateur de développement double de l'équation (8), pour $f = 2$ et $f = 3$ respectivement. En ce qui concerne l'estimateur de développement double, nous avons testé les répétitions décrites aux équations (16) et (17), que nous appellerons respectivement variantes 1 et 2.

Nous avons aussi établi l'estimateur de variance jackknife qui correspondait à l'estimateur intégral de la première phase pour chacun des $R = 4\ 000$ échantillons de la première phase, aux fins de comparaison. On obtient cet estimateur avec l'équation (8), quand $f = 1$.

Nous avons étudié diverses propriétés de fréquence des estimateurs précités et de l'estimateur de variance jackknife qui y correspond. Ces propriétés apparaissent ci-dessous. Pour plus de simplicité, elles ne sont exprimées qu'en fonction du nombre total estimatif de personnes occupées.

Le biais relatif en pour cent du nombre estimé de personnes occupées par rapport à la population globale est estimé par

$$\mathrm{BRP}(t^*) = \{[E_M(t^*)/T_y] - 1\} \times 100, \qquad (18)$$

où

$$E_M(t^*) = (1/4\ 000) \sum_{r=1}^{4\ 000} t_r^*$$

représente l'espérance de Monte Carlo de l'estimateur ponctuel t^* applicable aux 4 000 échantillons. La valeur t^* peut correspondre à t_1, t_2, ou t_3, alors que t_r^* est la valeur t^* de l'échantillon r.

où

$$r_{hji} = y_i - \sum_{k \in S_g} w_{hjk} y_k \Big/ \sum_{k \in S_g} w_{hjk} \quad \text{pour} \quad i \in s_g.$$

Sous réserve de légères conditions (voir les équations (A2) et (A3) de l'annexe), on obtient l'équation que voici, analogue à l'équation (5):

$$t_{(hj)2} \approx t_{(hj)1} + \sum_{g=1}^{G} (M_g/m_g) \sum_{i \in s_g} w_{hji} r_{hji}$$
$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + [M_g/m_g] c_i r_{hji}), \tag{9}$$

où c_i est une variable indicatrice égale à 1 quand i fait partie du sous-échantillon et a la valeur nulle dans les autres cas.

Poursuivons,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + \{[M_g/m_g] c_i - 1\} r_{hji})$$
$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_{hji}, \tag{10}$$

où $z_{hji} = y_i + \{[M_g/m_g] c_i - 1\} r_{hji}$. Une fois encore, puisque la valeur de m_g est toujours élevée, on peut raisonnablement supposer que $r_{hji} \approx r_i$ (voir l'équation (A4) de l'annexe). Par conséquent,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_i,$$

où $z_i = y_i + \{[M_g/m_g] c_i - 1\} r_i$. Pour les mêmes raisons, $t_2 \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_i z_i$. Puisque t_2 est linéaire dans z_i,

$$v_{J2} \approx v_{L1}(\textstyle\sum \sum w_i z_i) = \displaystyle\sum_{h=1}^{H} (n_h/[n_h - 1])$$
$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i z_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right]^2 \Big/ n_h \right). \tag{11}$$

Soit $e_i = M_g/m_g$, le facteur de pondération à la deuxième phase pour $i \in S_g$. On constate que c_i est une variable aléatoire, $E(c_i) = m_g/M_g$ et que $E(c_i c_k) = (m_g/M_g)(m_g - 1)/(M_g - 1)$ pour $i, k \in S_g, i \neq k$.

Il s'ensuit que

$$E_2\left[\left(\sum_{i \in U_{hj}} w_i z_i\right)^2\right] \approx \left(\sum_{i \in U_{hj}} w_i y_i\right)^2 + \sum_{i \in U_{hj}} (e_i - 1)(w_i r_i)^2$$
$$- \sum_{g=1}^{G} \sum_{\substack{i,k \in S_g \cap U_{hj} \\ i \neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \tag{12}$$

Pareillement, en appelant F_h^* l'ensemble des éléments venant des grappes tirées de la strate h du premier degré avant le sous-échantillonnage, on obtient

$$E_2\left[\left(\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i\right)^2\right] = E_2\left[\left(\sum_{i \in F_h^*} w_i z_i\right)^2\right]$$
$$\approx \left(\sum_{i \in F_h^*} w_i y_i\right)^2 + \sum_{i \in F_h^*} (e_i - 1)(w_i r_i)^2$$
$$- \sum_{g=1}^{G} \sum_{\substack{i,k \in S_g \cap F_h^* \\ i \neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \tag{13}$$

Dans l'annexe, on suppose que le dernier terme des équations (12) et (13) est négligeable, sous réserve de légères conditions. Par conséquent,

$$E_2(v_{J2}) \approx v_{J1} + \sum_{h=1}^{H} \sum_{i \in F_h^*} (e_i - 1)(w_i r_i)^2$$
$$= v_{J1} + \sum_{g=1}^{G} \sum_{i \in S_g} ([M_g/m_g] - 1)(w_i r_i)^2$$
$$\approx v_{L1} + E_2[(t_2 - t_1)^2], \tag{14}$$

qui, à son tour, implique que v_{J2} donne une estimation presque non biaisée de $E[(t_2 - T)^2]$.

4. L'ESTIMATEUR DE DÉVELOPPEMENT DOUBLE

L'estimateur de développement double est une solution de rechange à t_2, et se présente comme suit:

$$t_3 = \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i y_i. \tag{15}$$

La répétition jackknife de t_3 ne se définit pas clairement. Une simple possibilité serait

$$t_{(hj)3} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{hji} (M_g/m_g) y_i. \tag{16}$$

Une autre, qui se rapproche peut-être davantage d'une véritable «répétition», est

$$t^*_{(hj)3} = \sum_{g=1}^{G} \sum_{i \in s_g} w_{hji} (M_{ghj}/m_{ghj}) y_i, \tag{17}$$

où M_{ghj} représente le nombre d'éléments de l'échantillon de la première phase (plus exactement d'une grappe de l'échantillon de la première phase) qu'on trouve dans S_g mais *pas* dans U_{hj}. Parallèlement, m_{ghj} correspond au nombre d'éléments de l'échantillon de la deuxième phase qu'on trouve dans s_g mais *pas* dans U_{hj}. À partir de contre-exemples, nous verrons l'annexe, qu'aucune variante de la répétition ne donne d'estimateur de variance jackknife (v_{J3} de l'équation (8)) non biaisé de façon asymptotique, en général.

Par conséquent,

$$t_2 - t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i \left\{ \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} - \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i r_i}{\sum_{i \in s_g} w_i},$$

où

$$r_i = y_i - \sum_{k \in S_g} w_k y_k \Big/ \sum_{k \in S_g} w_k \text{ pour } i \in S_g.$$

Dans l'argumentation subséquente, il est capital de se rappeler que r_i a été défini afin que $\sum_{i \in S_g} w_i r_i = 0$ pour toutes les valeurs de g.

Si on poursuit,

$$t_2 - t_1 \approx \sum_{g=1}^{G} \sum_{i \in s_g} (M_g / m_g) w_i r_i, \tag{5}$$

puisque $\sum_{i \in S_g} w_i \approx \sum_{i \in s_g} (M_g / m_g) w_i$ (voir l'équation (A1) de l'annexe). Il s'ensuit que

$$E_2[(t_2 - t_1)^2] \approx \mathrm{Var}_2 \left\{ \sum_{g=1}^{G} \sum_{i \in s_g} (M_g / m_g) w_i r_i \right\}$$

$$= \sum_{g=1}^{G} (M_g^2 / [\{M_g - 1\} m_g])(1 - m_g / M_g)$$

$$* \left\{ \sum_{i \in S_g} (w_i r_i)^2 - \left(\sum_{i \in S_g} w_i r_i \right)^2 / M_g \right\}$$

$$\approx \sum_{g=1}^{G} ([M_g / m_g] - 1) \left\{ \sum_{i \in S_g} (w_i r_i)^2 \right\}. \tag{6}$$

Précisons que l'équation (6) *tient compte* des corrections pour la population finie qui résultent de l'échantillonnage de la deuxième phase.

3. L'ESTIMATEUR DE VARIANCE JACKKNIFE

3.1 L'estimateur de variance

Le moment est venu de parler de l'estimateur jackknife. Pour $j \in F_h$, soit la répétition $t_{(hj)2}$ de l'estimateur jackknife

$$t_{(hj)2} = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} \right\}, \tag{7}$$

où

$$w_{hji} = \begin{cases} w_i n_h / (n_h - 1) \text{ quand } i \in U_{hj'} \text{ et } j' \neq j \\ 0 \text{ quand } i \in U_{hj} \\ w_i \text{ quand } i \in U_{h'j'} \text{ et } h' \neq h. \end{cases}$$

De même, on peut dire que

$$t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} y_i.$$

Selon Rust (1985), l'*estimateur de variance jackknife* $v_{Jf} (f = 1 \text{ or } 2)$, se définit simplement par

$$v_{Jf} = \sum_{h=1}^{H} (n_h - 1)/n_h \sum_{j \in F_h} (t_{(hj)f} - t_f)^2. \tag{8}$$

Krewski et Rao (1981, équation (2.4)) dénotent cette forme $v_J^{(2)}$. On peut démontrer aisément que $v_{J1} = v_{L1}$.

3.2 Pourquoi l'estimateur fonctionne (un peu plus de théorie)

Nous verrons bientôt que v_{J2} estime presque sans biais la variance de l'estimateur de développement repondéré de l'équation (2). Rao et Shao (1992) parviennent indirectement à la même conclusion (notre équation (2) correspond à l'espérance de l'estimateur qu'ils présentent à la partie 3.3, pp. 818-819). Dans leurs travaux cependant, ces auteurs considèrent la non-réponse comme une phase d'échantillonnage supplémentaire où on recourt à l'échantillonnage de Poisson (Särndal et coll. 1992, p. 85) plutôt qu'à un échantillonnage aléatoire simple stratifié. Dans la démonstration de Rao et Shao, chaque élément prélevé à la première phase constitue en réalité une strate de deuxième phase. La quasi-absence de biais pour v_{J2} se résume donc au cas particulier d'un résultat signalé par Krewski et Rao (1981), (Rao et Shao (1992), p. 821).

Par «strate de deuxième phase», nous entendons les classes de repondération de Rao et Shao (1992). On présume que les éléments d'une classe donnée présentent la même probabilité inconnue de réponse ou de sélection. L'échantillonnage de Poisson équivaut à un échantillonnage aléatoire simple stratifié, *conditionnellement à* la taille du sous-échantillon obtenu à l'intérieur de la classe de repondération. Dans leurs travaux, Rao et Shao (1992) utilisent une approche *inconditionnelle*.

Revenant au problème qui nous intéresse, on remarque que

$$t_{(hj)2} - t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} \left\{ \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} - \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} r_{hji}}{\sum_{i \in s_g} w_{hji}} \right\},$$

variance jackknife ne présente presque aucun biais à l'égard de l'estimateur de variance repondéré, tandis que la partie 4 expose les lacunes de l'estimateur jackknife en tant qu'estimateur de la variance de l'estimateur de développement double. À la partie 5, on trouvera une étude de simulation qui semble confirmer les grandes hypothèses des parties antérieures. La partie 6 aborde les applications de l'estimateur de développement repondéré et la partie 7 sert de conclusion. L'annexe résume le cadre asymptotique hypothétique utilisé comme point de départ et fournit des éléments de preuve.

2. L'ESTIMATEUR DE DÉVELOPPEMENT REPONDÉRÉ

2.1 L'estimateur

Soit $h(=1, ..., H)$, les strates de la première phase d'un échantillon aléatoire en grappes stratifié, obtenu par tirage non exhaustif; n_h le nombre de grappes de la strate h échantillonnées et F_h l'ensemble des grappes. Soit $g(=1, ..., G)$, la strate de la deuxième phase d'où on prélève par tirage exhaustif un sous-échantillon aléatoire simple stratifié. Un élément de la grappe prélevé p fois lors de la première phase donne p éléments distincts pour le sous-échantillon. Soit M_g le nombre d'éléments dans g avant le sous-échantillonnage et m_g le nombre d'éléments sous-échantillonnés dans g. Dans la pratique, les strates de la deuxième phase G sont rarement définies avant prélèvement de l'échantillon de la première phase.

Soit S_g l'ensemble des éléments dans g avant le sous-échantillonnage; s_g le jeu d'éléments sous-échantillonnés dans g; s, l'ensemble complet d'éléments sous-échantillonnés et $m = \sum_g m_g$ la taille du sous-échantillon. Enfin, soit y_i la valeur à laquelle on s'intéresse pour l'élément i et w_i le facteur de développement à la première phase de i (c'est-à-dire, la valeur inverse de la probabilité de sélection de la grappe renfermant i).

En supposant le dénombrement de tous les éléments de l'échantillon de la première phase, pour estimer la population totale T, on recourrait à l'estimateur

$$t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i y_i. \tag{1}$$

Soit *l'estimateur de développement repondéré* de T,

$$\begin{aligned} t_2 &= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} (M_g/m_g) w_i y_i}{\sum_{i \in s_g} (M_g/m_g) w_i} \right\} \\ &= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} \right\}. \end{aligned} \tag{2}$$

Une autre façon d'écrire t_2 est

$$t_2 = \sum_{g=1}^{G} \sum_{i \in s_g} a_i y_i = \sum_{i \in s} a_i y_i, \tag{3}$$

où

$$a_i = \left[\sum_{k \in S_g} w_k \Big/ \sum_{k \in s_g} w_k \right] w_i \text{ pour } i \in s_g$$

correspond au *poids pondéré* de l'élément i. L'équation (3) explique le nom donné à l'estimateur (estimateur de développement repondéré).

2.2 Erreur quadratique moyenne de l'estimateur (un peu de théorie)

En général, t_2 donne une estimation biaisée de T. Toutefois, sous de légères conditions, précisées en annexe, l'estimateur de T est cohérent avec le plan d'échantillonnage. En d'autres termes, $\text{plim}_{m \to \infty} (t_2 - T)/T = 0$ (Isaki et Fuller 1982). Dans notre article, on supposera simplement que m_g est élevé.
Notons que

$$E[(t_2 - T)^2] = E[(\{t_1 - T\} + \{t_2 - t_1\})^2]$$

$$\approx \text{Var}_1(t_1) + E_1\{E_2[(t_2 - t_1)^2]\},$$

où les indices de Var et de E indiquent la phase de l'échantillonnage. Étant donné la valeur élevée de m_g, $E_2[t_1(t_2 - t_1)] = t_1 E_2(t_2 - t_1) \approx 0$. En outre, $E(t_2 - T) = E_1[E_2(t_2 - T)] \approx 0$, et l'erreur quadratique moyenne de t_2 correspond en réalité à sa variance (asymptotique).

Puisqu'on a procédé à un échantillonnage non exhaustif à la première phase, $\text{Var}_1(t_1)$ peut en principe être estimé au moyen de l'estimateur suivant:

$$v_{L1} = \sum_{h=1}^{H} (n_h/[n_h - 1]) * \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i y_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i y_i \right]^2 / n_h \right), \tag{4}$$

où U_{hj} correspond à l'ensemble des éléments de la grappe j tirée de la strate h à la première phase. L'indice L est utilisé pour des raisons historiques, pour indiquer qu'il y a «linéarisation», même s'il n'y a rien à linéariser dans le cas actuel. Notons que quand on effectue un deuxième échantillonnage, il s'avère généralement impossible de calculer v_{L1}, dans la pratique.

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 89-98
Statistique Canada

La méthode du jackknife convient-elle à un échantillon à deux phases?

PHILLIP S. KOTT et DIANA M. STUKEL[1]

RÉSUMÉ

L'estimateur de variance jackknife présente des propriétés intéressantes quand on s'en sert avec les estimateurs lisses tirés d'échantillons stratifiés à plusieurs degrés. L'article que voici porte sur l'application de cet estimateur à un plan particulier d'échantillonnage à deux phases: on commence par constituer un échantillon aléatoire stratifié en grappes par tirage non exhaustif, puis on restratifie les éléments des grappes échantillonnées et on prélève des sous-échantillons aléatoires simples de chaque strate de la deuxième phase. Apparemment, l'estimateur jackknife donne des résultats raisonnables pour ce qui est d'estimer la variance d'un estimateur de «développement» commun, mais pas celle d'un autre. Les auteurs parlent de l'application de leurs résultats à des stratégies d'estimation plus complexes. Une étude de Monte Carlo étaye leurs principales constatations.

MOTS CLÉS: Stratifié; estimateur de développement repondéré; estimateur de développement double; asymptotique.

1. INTRODUCTION

Krewski et Rao (1981) et, après eux, Rao et Wu (1985) se sont penchés sur les propriétés de plan de sondage de l'estimateur de variance jackknife dans le cas d'une stratification à plusieurs degrés intégrant un échantillonnage non exhaustif au premier degré. Bien qu'assez généraux en soi, les résultats obtenus par ces chercheurs ne peuvent directement être appliqués à bon nombre de plans d'échantillonnage à plusieurs phases. On lira à ce sujet Wolter (1985; chapitre 4.5).

Nous examinerons ici un simple exemple d'échantillonnage à deux phases. Dans un premier temps, on prélève un échantillon aléatoire stratifié en grappes, par tirage non exhaustif. Les éléments des grappes échantillonnées subissent ensuite une nouvelle stratification, peut-être au moyen de renseignements recueillis à la première phase, et on construit de façon aléatoire un nouveau sous-échantillon simple stratifié, par tirage exhaustif.

Il est possible d'estimer un total sans information auxiliaire de deux façons. La première consiste à multiplier la valeur de chaque élément sous-échantillonné par le produit de ses facteurs de pondération à chaque phase (à savoir, l'inverse de la probabilité de sélection à la première et à la deuxième phase), puis d'en faire la somme. C'est *l'estimateur de développement double* que Särndal, Swensson et Wretman (1992, p. 347) appellent «estimateur π^*».

Si on parle beaucoup de l'estimateur de développement double dans les traités de statistique, dans la pratique, il est plus courant de recourir à *l'estimateur de développement repondéré*, surtout si on traite la non-réponse des comme une deuxième phase de l'échantillonnage, comme Oh et Scheuren (1983, p. 150) le font avec l'estimateur de la classe de pondération. Pour obtenir un estimateur de la taille de la population applicable aux strates de la deuxième phase d'échantillonnage, on additionne les facteurs de pondération de la première phase pour tous les éléments de la strate avant le sous-échantillonnage. Ensuite, on multiplie le résultat par la moyenne estimée de la strate de la deuxième phase à partir du sous-échantillon, ce qui donne le total estimé de la strate. Enfin, on fait la somme des totaux estimés pour les strates de la deuxième phase de l'échantillonnage, ce qui produit l'estimateur de développement repondéré de la population totale.

Dans le cas présent, nous nous intéresserons plus à un véritable échantillonnage à deux phases qu'à l'usage de la non-réponse comme phase d'échantillonnage artificielle supplémentaire. Le National Agricultural Statistics Service (NASS) recourt actuellement à l'estimateur de développement double dans ses enquêtes trimestrielles sur l'agriculture (ETA). On obtient un échantillon stratifié, ariolaire en grappe est de nombre en juin. Les exploitations agricoles identifiées en juin sont restratifiées d'après les réponses données le même mois, puis rééchantillonnées en vue d'un nouveau dénombrement en septembre, en décembre et en mars.

Le NASS se sert d'un plan d'échantillonnage à deux phases et de l'estimateur de développement repondéré dans le cadre de son enquête sur l'usage des produits agrochimiques à la ferme. On commence par identifier les exploitations qui produisent certaines cultures, puis on quantifie l'emploi de pesticides avec ces cultures.

Le présent article montre que si on peut utiliser l'estimateur jackknife pour estimer la variance de l'estimateur de développement repondéré dans certaines conditions, cette méthode ne s'avère pas très efficace, en général, pour estimer la variance de l'estimateur de développement double. À la partie 2, il est question de l'estimateur de développement repondéré et son erreur quadratique moyenne. À la partie 3, on verra que l'estimateur de

[1] Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030; Diana M. Stukel, Division des méthodes d'enquête des ménages, Statistique Canada, Ottawa, (Ontario), Canada, K1A OT6.

Singh, Tsui, Suchindran et Narayana expliquent le plan d'enquête et les techniques d'estimation auxquels on a recouru dans le cadre de PERFORM (examen de l'évaluation des projets en vue de la gestion des ressources organisationnelles), enquête de grande envergure qui s'est déroulée dans l'État d'Uttar Pradesh, en Inde, qui devait servir à estimer les caractéristiques des installations de santé et de la population desservie, de manière à établir les valeurs repères essentielles à un important projet de planification familiale. PERFORM fait appel à un plan d'échantillonnage stratifié à degrés multiples avec pour unités d'échantillonnage les ménages et les femmes admissibles qui en sont membres. On estime toutefois aussi les services de santé qui ne se retrouvent pas explicitement dans le plan d'échantillonnage en procédant à une correction qui tient compte de la multiplicité des unités d'échantillonnage secondaires sélectionnées, auxquelles les installations de santé procurent leurs services.

Dufour, Kaushal et Michaud passent en revue les tests et les études qui ont précédé l'application de l'interview assistée par ordinateur à la plupart des enquêtes-ménages, à Statistique Canada. L'interview se donne en personne, au domicile du répondant, ou au téléphone, du domicile de l'intervieweur, grâce à un ordinateur portatif. Les auteurs parlent des difficultés qu'a soulevées l'implantation de cette nouvelle technologie au niveau des enquêtes permanentes et des nouvelles possibilités qu'elle laisse entrevoir quant au contrôle de la collecte des données.

Scheuren et Winkler proposent une méthode autorisant l'emploi de variables quantitatives peu courantes mais corrélées en vue d'améliorer le couplage des enregistrements. L'idée fondamentale consiste à recourir aux couplages dont l'exactitude est presque assurée pour estimer le lien entre les variables peu courantes par régression et utiliser les valeurs prévues des mêmes variables lors d'un second couplage des enregistrements. On peut reprendre cette méthode par itération jusqu'à ce qu'il y ait convergence. La régression fait appel à une technique où les valeurs de régression sont corrigées des erreurs que pourrait présenter le couplage, ainsi qu'on a déjà pu le lire dans un article des mêmes auteurs, publié dans le numéro de juin 1993 de *Techniques d'enquête*. Après illustration empirique, on montre que cette méthode peut déboucher sur de bons résultats dans des situations qui paraissaient jusqu'alors sans issue.

Le rédacteur en chef

Cher(ère) lecteur(trice) de *Techniques d'enquête,*

J'aimerais profiter de cette occasion pour vous remercier de l'intérêt et de l'appui manifesté à la publication *Techniques d'enquête*. Depuis sa création, cette revue publie des articles qui intéressent les organismes statistiques et les checheurs(euses) en accordant une attention particulière à l'élaboration et à l'évaluation de techniques précises appliquées à la collecte des données ou aux données elles-mêmes.

La revue *Techniques d'enquête* célébrera bientôt son 25$^{\text{ième}}$ anniversaire. Depuis son début en tant que revue interne des développements de méthodologie d'enquête à Statistique Canada, elle a évolué en une revue statistique largement consultée avec un comité de rédaction de statisticiens reconnus à travers le monde. Bien que de nombreuses modifications y aient été apportées en vue d'en améliorer le contenu et la présentation, il y a toujours matière à amélioration. Ainsi, je vous invite à nous faire part de tous commentaires, suggestions ou recommandations susceptibles de nous aider à continuer de faire de *Techniques d'enquête* une plate-forme fiable du développement des statistiques du prochain millénaire.

Si vous désirez qu'un exemplaire de *Techniques d'enquête* soit envoyé à titre gracieux à un collègue, n'hésitez pas à communiquer avec nous.

En terminant, j'aimerais à nouveau vous remercier de votre intérêt et appui à notre revue *Techniques d'enquête.*

Je vous prie d'agréer, Monsieur, Madame, l'expression de mes sentiments les meilleurs.

M.P. Singh
singhmp@statcan.ca

Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 87-88
Statistique Canada

Dans ce numéro

Le numéro de *Techniques d'enquête* que voici renferme des articles sur des sujets variés. Kott et Stukel étudient l'estimation de la variance jackknife pour un plan d'échantillonnage à deux degrés particulier, mais d'un vaste usage. En un premier temps, on sélectionne des grappes dans les strates par échantillonnage aléatoire simple avec remise et on retient tous les sujets des grappes sélectionnées. À la deuxième étape, les sujets échantillonnés font l'objet d'une nouvelle stratification et un échantillonnage aléatoire simple permet d'obtenir les unités du deuxième degré. Les auteurs examinent deux estimateurs ponctuels: «l'estimateur d'expansion repondéré» et «l'estimateur d'expansion double», plus courant. Avec un tel plan d'échantillonnage, on constate que l'estimateur de la variance jackknife se comporte étonnamment mieux avec le premier estimateur ponctuel qu'avec le second. Une étude de Monte Carlo confirme cette constatation.

Decaudin et Labat présentent un système d'estimation de population «multi-sources» visant à produire des estimations locales de population durant les périodes intercensitaires en France. Le système présenté est robuste et souple en ce qu'il fonctionne avec un nombre variable de sources. Il repose sur une synthèse robuste d'estimations provenant de différentes sources, en combinant un raisonnement démographique et des techniques statistiques.

Ravalet applique les GM-estimateurs avec une procédure adaptative à l'enquête sur l'investissement industriel de l'INSEE, afin de produire un estimateur robuste. Les fonctions examinées sont la fonction bicarrée de Tukey et la fonction de Cauchy. Chacune de ces deux fonctions dépend d'une constante de réglage qui est choisie en fonction de l'épaisseur de queue de la distribution des observations et de la concentration des résidus. Les constantes de réglage qui minimisent la variance de l'estimateur sont trouvées pour huit distibutions particulières présentant diverses situations quant à l'épaisseur de queue et la concentration des résidus supposés symétriques.

Cotton et Hesse examinent les caractéristiques de plusieurs méthodes de sélection d'un panel stratifié de taille fixe, et leur impact sur la sélection intitiale, la rotation, le retirage ainsi que le recouvrement de l'échantillon. Les auteurs proposent un type d'algorithme basé sur des transformations des numéros aléatoires permanents servant aux tirages qui prolonge après retirage la rotation avant retirage. Ces transformations peuvent être effectuées sur les numéros aléatoires rendus équidistants, ainsi que sur les numéros aléatoires provenant d'une loi uniforme.

Dans son article, Farrell étudie l'estimation empirique de Bayes pour des proportions de petites régions. Les données du recensement américain lui permettent de comparer les estimations bayésiennes de petites régions de la proportion de personnes se retrouvant dans telle ou telle tranche de revenu obtenues de façon empirique au moyen de modèles logistiques multinomiaux et ordinaux avec effets aléatoires. Les inférences issues du modèle ordinal sont légèrement préférables à celles du modèle multinomial. L'auteur compare aussi les estimations rajustées de la variance venant des modèles naïf et «bootstrap», de même que la probabilité de couverture des intervalles de confiance qui s'y associent. La correction obtenue par la méthode «bootstrap» améliore sensiblement la couverture.

Gelman et Little décrivent un nouveau prolongement de l'analyse des données d'enquête stratifiées a posteriori faisant appel à une modélisation bayésienne par régression logistique hiérarchique. Cette technique engendre beaucoup plus de catégories de stratification qu'on en obtient typiquement avec les méthodes habituelles de stratification a posteriori et de pondération, si bien que le modèle peut englober une somme beaucoup plus grande d'informations au niveau de la population. Les auteurs appliquent la méthode qu'ils proposent et d'autres méthodes plus classiques aux données des sondages d'opinion précédant les élections aux États-Unis avant de procéder à une évaluation graphique des divers modèles en comparant leurs résultats à l'issue véritable des élections.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 23, numéro 2, décembre 1997

TABLE DES MATIÈRES

Dans ce numéro 87

P.S. KOTT et D.M. STUKEL
La méthode du jackknife convient-elle à un échantillon à deux degrés? 89

G. DECAUDIN et J.-C. LABAT
Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population en France 99

P. RAVALET
Une procédure adaptative d'estimation robuste du taux d'évolution de l'investissement 107

F. COTTON et C. HESSE
Tirage et maintenance d'un panel stratifié de taille fixe 117

P.J. FARRELL
Estimation de proportions pour petites régions par des méthodes empiriques de Bayes, à partir de variables ordinales 127

A. GELMAN et T.C. LITTLE
Stratification a posteriori en un grand nombre de catégories par régression logistique hiérarchique 135

K.K. SINGH, A.O. TSUI, C.M. SUCHINDRAN et G. NARAYANA
Estimation de la population et des caractéristiques des établissements de santé et des populations de clients au moyen d'un plan d'échantillonnage à plusieurs degrés avec enchaînement 147

J. DUFOUR, R. KAUSHAL et S. MICHAUD
Les interviews assistées par ordinateur dans un environnement décentralisé: Le cas des enquêtes-ménages à Statistique Canada 159

F. SCHEUREN et W.E. WINKLER
Analyse de régression des fichiers de données appariés par ordinateur - Partie II 171

Remerciements 181

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (nº 12-001-XPB au catalogue) est de 47 $ par année au Canada et de 47 $ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 1997 • VOLUME 23 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

Février 1998

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

4427

TECHNIQUES D'ENQUÊTE

Nº 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

DÉCEMBRE 1997

•

VOLUME 23

•

NUMÉRO 2

Statistique Canada   Statistics Canada

Canada